PERSPECTIVE

nature genetics

Developing predictive molecular maps of human disease through community-based modeling

Jonathan M J Derry, Lara M Mangravite, Christine Suver, Matthew D Furia, David Henderson, Xavier Schildwachter, Brian Bot, Jonathan Izant, Solveig K Sieberts, Michael R Kellen & Stephen H Friend

The inability to identify the molecular causes of disease has led to a disappointing rate of development of new medicines. By combining the power of community-based modeling with broad access to large datasets on a platform that promotes reproducible analyses, we can work toward more predictive molecular maps that can deliver better therapeutics.

A better molecular understanding of disease is needed

Drugs continue to fail in clinical development at a startlingly high rate despite unprecedented amounts of investment in research and development, largely as a result of a lack of efficacy in phase 2 trials[1]. This lack of efficiency stems from a failure in biology in selecting the correct target rather than a chemical failure; many compounds are shown to be safe and to engage the intended target, but they do not improve the primary clinical indication. This breakdown has its origins in the simplistic ways in which we identify potential drug targets for complex diseases and indicates a need for more innovative approaches to identify causal relationships between molecular entities and disease.

Biology is rapidly becoming a science that is driven by technology and large-scale data. Herein lies an opportunity to transform our understanding of the molecular underpinnings of disease and develop modeling frameworks that can describe complex systems and predict their behavior. At one level, a simple pairwise analysis of alterations in human diseases may be useful for providing lists of altered components, but to uncover the essential mechanistic relationships between molecular changes and disease, more integrative modeling methods that combine multiple complex molecular traits with phenotypic outcomes will be required[2–5]. It is probable that the particular approach used will be linked to the question being addressed, such that problems of classification—for example, for disease outcome or drug response—may require different models from those used for questions directed at understanding mechanisms and predicting therapeutic intervention points.

Sage Bionetworks, Seattle, Washington, USA. Correspondence should be addressed to J.M.J.D. (derry@sagebase.org) or S.H.F. (friend@sagebase.org).

Published online 27 JANUARY 2012; doi:10.1038/ng.1089

Building a Commons as a means to develop maps of disease

The challenge of generating predictive molecular models of disease is complex and is not likely to be solved by any one group of researchers. Instead, it will be necessary for researchers in the field of biology to adopt the community-based practices that have proven successful in other areas of science and technology. Enabling scientists to reproduce and extend the work of others will require that the data and methods used be distributed in a manner that is both accessible and usable. Despite efforts by funding agencies and publishers, data sharing is intermittent, and data that are made accessible are often done so in a way that does not provide sufficient information for the reuse of the data. In part, this problem stems from the current lack of a suitable mechanism for ensuring reproducibility of data and analyses, with print journals being a poor avenue for hosting large datasets and complex algorithms[6]. Without the provision of sufficient methodological detail and direct access to the data, as well as the code and workflows used to produce the particular analyses, the results of modeling approaches are not broadly useful to the community and do not advance biological understanding[7].

We advocate the concept of a 'Commons', in which contributor scientists can collaborate in transparent and structured ways to build better maps of disease from a common reference of curated data. In this vision, the contributors are not simply people who upload or download data for isolated use but, instead, they are active participants that build collective content in a manner analogous to other distributed community projects, such as Wikipedia. Such a system could effectively crowdsource the evolution of better disease models and would provide an accelerated mechanism for the dissemination of knowledge. In this Perspective, we describe key aspects of the Sage Bionetworks Commons project, including the efforts made to date in building a computational platform and a data and model repository that includes the associated analysis tools, as well as the development of data sharing rules and policies. We explain how this environment will drive us toward the generation of better maps of disease and become a forum for reproducible and reusable data and analyses. Community involvement will be necessary to address the many concerns that such a complex endeavor will encounter, including ways to incentivize data sharing, to promote the appropriate attribution for the data generators and map builders and to address policy issues associated with the protection of human data. Engagement of stakeholders across different constituencies to drive the development of the policies and resources necessary for the project is crucial (http://sagebase.org/WP/com/?page_id=14).

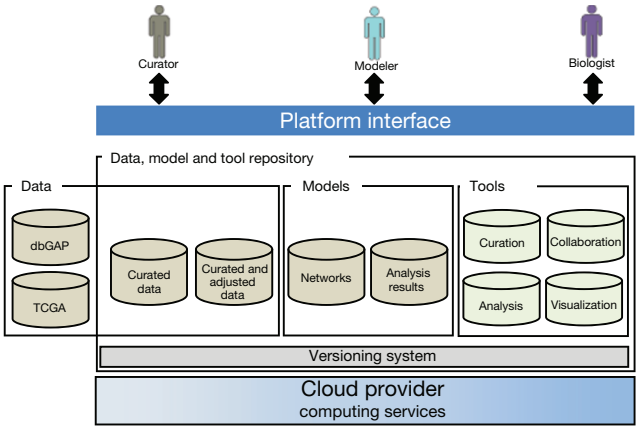

Figure 1 Synapse platform architecture. Synapse uses a set of web services to provide access to the data repository, which comprises a federated collection of curated, adjusted and analyzed datasets, models and code. Synapse may also reference restricted data stored in external databases, such as dbGAP or The Cancer Genome Atlas (TCGA). All resources managed by Synapse can be referenced as objects using a URL according to linked data principles. This approach allows for the storage of data and metadata using persistence mechanisms that are appropriate for each data modality while abstracting clients away from the details of how data and services are obtained. Integration with ontology services and support for a rich query language occurs on the Synapse backend, allowing multiple clients (for example, R and the web client) to run similar queries across hosted data. Versioning of data, workflows and tools allows for the documentation of details on how individual models were generated, and enables these models to be reproduced. Storage of the data repository and services in the cloud allows for scalability, access and the potential to use high performance computing facilities directly from Synapse.

The ultimate goal of this project is to provide a mechanism for the collaborative generation, modification and improvement of predictive computational models of disease. The current standard of publishing modeling methods as general descriptions in manuscripts does not provide sufficient detail for an analysis to be accurately reproduced. Distribution of analyses, linked to underlying data and detailed and versioned code, as well as analytical workflows, will ensure that models are built in a reproducible manner. This transparency will encourage collaborative analyses, provide a mechanism for the meaningful assessment of analytical quality and provide a forum for the development of analytical standards. Ultimately, this process will guide biological researchers to high quality analytical results from which they can inform their own research efforts.

Synapse is a platform for collaborative research

Central to the Commons is 'Synapse', a platform resource that enables a community-based genomic analysis and provides broad access to molecular models of disease and the underlying datasets and algorithms that were used to construct them (http://synapse.sagebase.org). Synapse provides various functionalities: the management of datasets, analysis code and models in user-created projects; the ability to publish these resources for public reuse; inclusion of a workflow and versioning system to track the specific dataset and code that was used for a particular analysis; and access to tools to enable scientific analysis and collaboration (**Fig. 1**). Synapse takes advantage of the maturing set of cloud computing technologies and will provide scientists with access to on-demand super-computer power without the upfront capital costs of building and managing a private cluster.

Synapse is built around a set of web services that provide a variety of features, including annotation, indexing, history tracking, versioning, authentication, authorization and data persistence. We designed an application-programming interface (API) that allows for structured queries across the metadata of all datasets, models and tools. Structured queries can be semantically enhanced by having Synapse delegate to external services, such as the National Center for Biomedical Ontology (NCBO)[8]. The API provides federated access to datasets and other objects managed by Synapse, which allows a single API to be used to query and load data regardless of whether the data are hosted directly on Synapse or are linked to an external system. This strategy allows the support of use cases where the data generator has imposed restrictions on its redistribution (for example, in dbGAP) or when data volumes are sufficient to preclude hosting in the public cloud. The API also allows analysis code to be brought to the data that is located with the computing resources. This obviates the need to download large data sets, a feature that will increasingly become a priority as genetic data volumes outstrip network transfer capacities.

The data, code and analytical results in Synapse are stored in a centralized repository (**Fig. 2**). With Synapse's versioning and provenance features, data and analyses can be stored and tracked from raw formats (for example, CEL files) through curation and quality control to analytical results. We aim to standardize data curation using specific software tools to facilitate both the preparation of the basic curated data files as well as their conversion to other data formats for downstream analysis (for example, using Bioconductor, MATLAB, PLINK or ISA-Tab)[9]. Though not a requirement for the data in Synapse, we recommend that data be converted to 'SageBio Curated' format, as the use of a non-proprietary standardized data format promotes interoperability across analyses. The process of data curation involves both data integrity checks and the transformation of the dataset into this standard text-based format. Curated datasets are the building blocks for shared and versioned analyses. A detailed description of the curation process and format are available at http://sagebase.org/commons/repository.php.

Synapse will also provide an 'analysis ready' version of each dataset by running the curated data through a quality control process in preparation for downstream analysis. These adjusted datasets will exist in conjunction with the source code and the detailed documentation that describes the transformation from the underlying curated and/or raw data (**Fig. 2**). In addition, Synapse will provide normalized versions of gene expression data from public databases (for example, Gene Expression Omnibus or ArrayExpress), as well as clinical phenotypes curated to existing ontologies, such as NCBO[8]. We recognize that normalization and data adjustment processes differ depending on the analytical goals, and so we anticipate that end users will create and store different versions of each dataset with the code for alternative adjustment strategies. Sage Bionetworks seeds the repository with versions of the data before and after quality control that are useful for modeling analyses, but over time, we expect the community of scientists to both deposit data and participate in the process of curation and quality control.

A key feature of Synapse will be access to the tools developed by the scientific community for the manipulation and analysis of data (**Fig. 2**). Synapse will support integration with applications that support various users, from data curators to bioinformaticians and biologists, and tools for data adjustment, normalization and reformatting, as well as for model building, will be developed and shared. For example, we have developed an R client to allow platform-hosted data to be accessed from the R environment, thereby providing a ready link to a wealth of existing analysis methodology (contained in the the Comprehensive

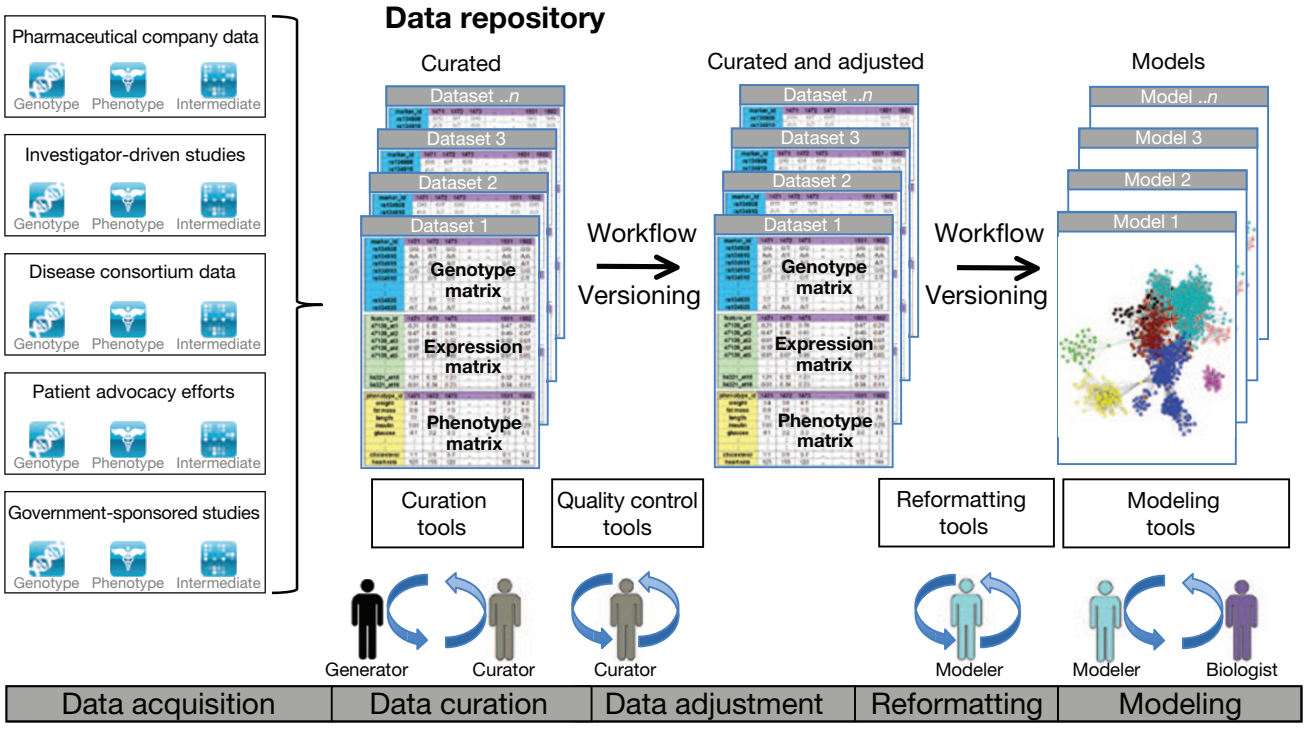

Figure 2 The process of data acquisition, curation, adjustment, reformatting and modeling. Data flows into the repository from a number of different sources (examples are shown). Individual datasets typically contain different types of data and are submitted in various formats. Curation involves reformatting the data into a common tab-delimited text matrix format. This curated standard format is available for download and allows for the development of workflows for common manipulations (for example, adjustments for technical covariates, such as gene expression array batch). The 'curated and adjusted' dataset is also available for download. Data analysts or modelers may use the curated data or the curated and adjusted data for downstream analyses; the key feature is that the version of the dataset that is used for an analysis, as well as the underlying code and workflow, is stored. Allowing different types of users to interact with the data at different points in the process has advantages. For example, providing tools to enable the curation of a dataset into a standard format provides the user with the benefit of easy curation and opens up tools for downsteam quality control and analysis.

R Archive Network (CRAN) and Bioconductor). Synapse also consists of a web portal that allows researchers to search and navigate through content relevant to their research interests and form projects with existing or new colleagues. General-purpose tools such as wikis, user forums and issue trackers can easily be adopted from other domains to support scientific research teams.

Data sharing and reproducible science

The benefits of sharing scientific data are widely acknowledged[10–15], serving to maximize the impact of research and accelerate the rate of scientific discovery, however, the implementation of processes aimed at broad sharing of data has lagged for multiple technical and cultural reasons. Appropriate recognition for the data-generating organization, as well as for those individuals involved in managing and curating the data, is a prerequisite for widespread data sharing. However, the funding and priorities for these activities are often unclear. For this reason, available data are often difficult to interpret or lack essential elements that are required for reuse. Mechanisms for the appropriate recognition of data generation and curation efforts are essential. Journal citation will be necessary, and further citation will need to use mechanisms for attribution beyond standard print publications[16]. Recognition of data sharing efforts within professional merit systems, including qualifications for tenure advancement, is also desirable. Some solutions are already being driven by publishers and funders through policies that require data deposition before manuscript publication and/or the completion of a funding cycle[17].

Sharing of human genomic data presents additional challenges. Common concerns include the maintenance of participant privacy and the risk that data misuse could lead to stigmatization or discrimination. To reduce such risks to participants and in accordance with federal and state laws that protect identifiable health information, all data available through Synapse must be stripped of direct Health Insurance Portability and Accountability Act (HIPAA) identifiers (http://www.hhs.gov/ocr/privacy/index.html). Data must be used with respect for the values and intentions of the study participants and in a manner that limits the risk for misuse. The community of data users and contributors must be accountable and act as stewards of the data. In this paradigm, shared expectations and trust among community participants is essential[18]. To this end, access to Synapse will be granted to registered users after authentication and agreement to standard terms and conditions of use, including attribution of data source, ethical use of data and the agreement not to attempt to re-identify human participants (http://sagebase.org/downloads/SageBio_TermsOfUse.pdf). Data submitters may indicate additional terms and use restrictions for certain datasets in accordance with informed consent directives or for datasets whose full disclosure could carry risks for participant privacy and/or group stigmatization.

Data users will be expected to fully comply with the limitations and restrictions set by the data submitter. Sage Bionetworks will never impose arbitrary restrictions to data use but, instead, will abide by restrictions outlined by the data contributor based on informed consent or guidance provided by the relevant institutional review board. The consequence of violating these rules is the denial of continued access to Synapse. A

public forum will be used to promote ethical behavior and prevent the misuse of the data in Synapse. Synapse will provide a way to give feedback and log concerns so that issues, whether logistic, scientific, ethical or regulatory, can be brought to the attention of the community and be promptly rectified. A number of efforts have aimed to develop principles and codes of conduct for the international sharing of genomic data (http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html)[19,20], and Synapse will follow these principles (http://sagebase.org/downloads/SageBio_Governance.pdf). In addition, efforts to change the participant consent process to provide greater control by participants regarding how and with whom their data are shared will also be essential to guarantee an individual's rights to selective data disclosure within an open source analytical environment[21].

Future directions

Synapse is designed to confer broad benefits across the biomedical research community, both to computational scientists who use analytical processes to generate new molecular models and to researchers who want to use these models to inform their own work in disease biology. This environment is designed to foster the development of more reliable models through iterative community improvements in analytical methodologies. Through the broader context of the Commons, Synapse will provide a mechanism to link model generators with researchers and clinicians poised to validate modeling hypotheses and incorporate modeling results into research directed at understanding physiological or disease states and therapeutic development efforts.

ACKNOWLEDGMENTS

This project is supported by program grant 3104672 from the Washington Life Sciences Discovery fund, grant U54CA149237 from the Integrative Cancer Biology Program of the National Cancer Institute and a grant to the Stanford Center for Biomedical Ontology (U54 HG004028).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates. *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
2. Barabási, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
3. Friend, S.H. The need for precompetitive integrative bionetwork disease model building. *Clin. Pharmacol. Ther.* **87**, 536–539 (2010).
4. Schadt, E.E., Friend, S.H. & Shaywitz, D.A. A network view of disease and compound screening. *Nat. Rev. Drug Discov.* **8**, 286–295 (2009).
5. Tegnér, J.N. *et al.* Computational disease modeling - fact or fiction? *BMC Syst. Biol.* **3**, 56 (2009).
6. Mesirov, J.P. Computer science. Accessible reproducible research. *Science* **327**, 415–416 (2010).
7. Gentleman, R. Reproducible research: a bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.* **4**, Article2 (2005).
8. Musen, M.A. *et al.* The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* published online, doi:10.1136/amiajnl-2011–000523 (10 November 2011).
9. Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**, 2354–2356 (2010).
10. Friend, S.H. Something in common. *Sci. Transl. Med.* **2**, 40ed6 (2010).
11. Hrynaszkiewicz, I. The need and drive for open data in biomedical publishing. *Serials* **24**, 31–37 (2011).
12. Schofield, P.N. *et al.* Post-publication sharing of data and tools. *Nature* **461**, 171–173 (2009).
13. Guttmacher, A.E., Nabel, E.G. & Collins, F.S. Why data-sharing policies matter. *Proc. Natl. Acad. Sci. USA* **106**, 16894 (2009).
14. Field, D. *et al.* Megascience. 'Omics data sharing. *Science* **326**, 234–236 (2009).
15. Science Staff. Dealing with data. Challenges and opportunities. Introduction. *Science* **331**, 692–693 (2011).
16. Giardine, B. *et al.* Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nat. Genet.* **43**, 295–301 (2011).
17. Anonymous. Data's shameful neglect. *Nature* **461**, 145 (2009).
18. Anderson, N. & Edwards, K. Building a chain of trust: using policy and practice to enhance trustworthy clinical data discovery and sharing. *Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies* 15–20 (ACM Press, New York, New York, USA, 2010).
19. Walport, M. & Brest, P. Sharing research data to improve public health. *Lancet* **377**, 537–539 (2011).
20. Knoppers, B.M. *et al.* Towards a data sharing Code of Conduct for international genomic research. *Genome Med.* **3**, 46 (2011).
21. Shelton, R.H. Electronic consent channels: preserving patient privacy without handcuffing researchers. *Sci. Transl. Med.* **3**, 69cm4 (2011).