Developing Predictive Molecular Maps of Human Disease through Community-based Modeling

Jonathan M.J. Derry, Lara M. Mangravite, Christine Suver, Matt Furia, David Henderson, Xavier Schildwachter, Jonathan Izant, Solveig K. Sieberts, Michael R. Kellen, Stephen H. Friend

Sage Bionetworks, 1100 Fairview Ave North, Seattle, WA 98109-1024, USA

Direct correspondence to:
Jonathan Derry: derry@sagebase.org
Stephen Friend: friend@sagebase.org

Perspective: A Need for Better Maps of Disease

The revolution in health care that was anticipated by the sequencing of the human genome has failed to materialize[1]. The failure rate for drugs in clinical development is still startlingly high despite unprecedented investment in R&D that reached a record $65 billion in 2009. This is largely due to the very high attrition rate for compounds in clinical development due to lack of efficacy in phase II trials: the success rate for progression of compounds from first-in-man studies to approved drug is only 11%[2]. This represents a failure of biology in selecting the correct target rather than a failure of chemistry; many compounds are shown to be safe and to engage the intended target but do not improve the primary indication. This failure stems from the simplistic ways in which we have historically studied 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 changing and becoming a technology and data-intensive science with the development of new instrumentation to measure various molecular states in greater detail. Herein lays 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. Without these models acting as maps, biologists risk drowning in an ever-growing sea of data. This vision for biology, to use large-scale data to model disease, reflects parallel developments in other scientific disciplines: for example, modeling future trends in climate based on complex meteorological information in atmospheric science. The term fourth paradigm has been coined for this “data intensive” science discovery to distinguish it from empiric, theoretical and computational approaches[3].

While the physical sciences have been dominated by deductive modeling approaches, these techniques have had a more scattered impact across biology[4]. Molecular dynamic simulation of protein structures and physiological systems models provide areas where biological systems can be well described using deductive models and validated using experimental data. However, when confronted by the enormous complexity and general lack of specific knowledge in biological systems these models are often too under-constrained to be applied. In these cases, methods based on statistical inference are an alternative way to identify key molecular relationships linked to disease phenotypes. A number of successes using this approach have been reported in the past several years[5-13]. At one level simple pair-wise analysis of alterations in human diseases, be it from DNA to phenotype in genome-wide association studies or from mRNA to phenotype in gene expression profiling studies, may be useful in providing essential lists of altered components. However, 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[14,15]. In the long-term, different modeling approaches will likely be necessary to deliver complete, detailed and accurate models that can predict the behavior of biological systems. The particular approach will be linked to the question being addressed such that problems of classification - for disease outcome, drug response, etc – may require different models from those directed at understanding mechanisms and predicting therapeutic intervention points.

Building a Commons as a Means to Develop Better Maps of Disease

The challenge of generating predictive molecular maps of disease is large and complex and is not likely to be solved by any one group. Instead, it will be necessary for biology to adopt the community-based practices that have proven successful in other areas of science and technology. At their core these efforts involve the open release of data for broad access. A good example for this comes from the Earth Sciences. The National Center for Atmospheric Research (NCAR) provides a "Community Data Portal" to over 8,000 datasets relevant to the atmosphere, the Earth system and the Sun[16]. In addition, it provides tools for the analysis of this data and hosts community developed models thereby providing national and regional decision-makers with the most advanced science in weather and climate modeling. Science can also learn from the commercial world how broad sharing of data can drive improvements in models: Netflix has been able to improve its movie recommendation engine by creating a public challenge seeded with data describing user ratings of movies[17]. In both cases, the release of large, well-curated datasets into the public domain has driven significant improvements in modeling techniques. Movements to similarly share data relevant for research into human disease mechanisms are beginning to gain broad attention[18-22].

To make data truly easy to use, and also to enable scientists to reproduce and build on the work of others, it is not simply enough to deposit it into the public domain. It is also necessary to curate and document the data as well as the methods, tools and workflows used in analysis. Once again, biology would do well to learn from examples in other disciplines, such as the open source software industry where successful projects provide not just accessible code, but also invest effort in supporting documentation and developer tools that help new engineers build on previous work. Biology has traditionally operated as a “single use science” in which the results of one group are not readily available to be built upon by others because of issues around access to data and methods and inadequate documentation of what was actually done. Despite efforts by funding agencies and publishers, data sharing is still intermittent and data that are made accessible are often done so in a fashion that does not provide sufficient information for data re-use. In part this stems from the current lack of a suitable mechanism for ensuring reproducibility, with print journals being a poor avenue for hosting large datasets and complex algorithms[23]. Without provision of sufficient methodological detail, the results of modeling analyses are not of use to the community and do not advance biological understanding[24]. Indeed why should it be possible to publish in high profile journals if other scientists cannot reproduce the analyses of the authors let alone build upon the results? As previously stated “an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data that produced the result”[25].

The concept of a “Commons” in which contributor scientists can collaborate together in transparent and structured ways to build better maps of disease from a common reference of curated data is central to an effort being driven by Sage Bionetworks. The goal is to develop an innovation space where scientists can compute and collaborate on various clinical and genomic data, use tools to build models of human disease and modify those built by others. In this vision, contributors are not simply people who upload or download data for isolated use as for conventional database efforts. Instead contributors are active participants that build collective content in a manner analogous to other distributed community projects such as Wikipedia. This uncouples the data generators from the data analyzers, effectively crowdsources the evolution of disease models, and provides an accelerated mechanism for the dissemination of knowledge.

Community involvement will be necessary to address the many concerns such a complex project will encounter, including ways to incentivize data sharing, attribution for data generators and map builders, and policy issues associated with human data protection. Engagement of stakeholders across different constituencies to drive the development of the policies and resources for the project has been important from the beginning and continues to be so[26]. In this commentary we will describe important aspects of the project including the efforts to date in building a platform and data and network model repository and the associated tools, the development of data-sharing rules and policies, and how this will drive us towards better maps of disease and a forum for reproducible and re-usable data and analyses.

The Sage Bionetworks Platform

Sage Bionetworks is a non-profit organization with a mission to create a Commons where integrative bionetworks evolved by contributor scientists accelerate the elimination of human disease (www.sagebase.org). Central to the Commons is the Sage Bionetworks Platform: a resource to provide broad access to molecular network models of disease, and the underlying datasets and algorithms used to construct them. For clarity throughout this commentary we will refer to the "Platform" as representing the infrastructure being developed to enable community-based genomic analysis. This includes the data and network repository, the IT infrastructure in which data are housed and accessed by users, and the tools that allow manipulation and analysis of the data (Figure 1).

IT Infrastructure: The Platform provides a number of functionalities: management of datasets, analysis code, and network models; a workflow and versioning system that can track the specific dataset and code that was used for a particular analysis; and a suite of tools to enable scientific analysis and collaboration (Figure 1). Datasets may be hosted by the Platform or linked to when the data generator or other groups hosting the data have imposed restrictions on its redistribution (e.g. dbGAP or TCGA). In cases where redistribution of data is not allowed the Platform will reference the data source and provide code necessary to process the data into a standard format once the user has received the data directly from the data source. Regardless of physical location, the federated collection of curated, adjusted, and analyzed datasets,

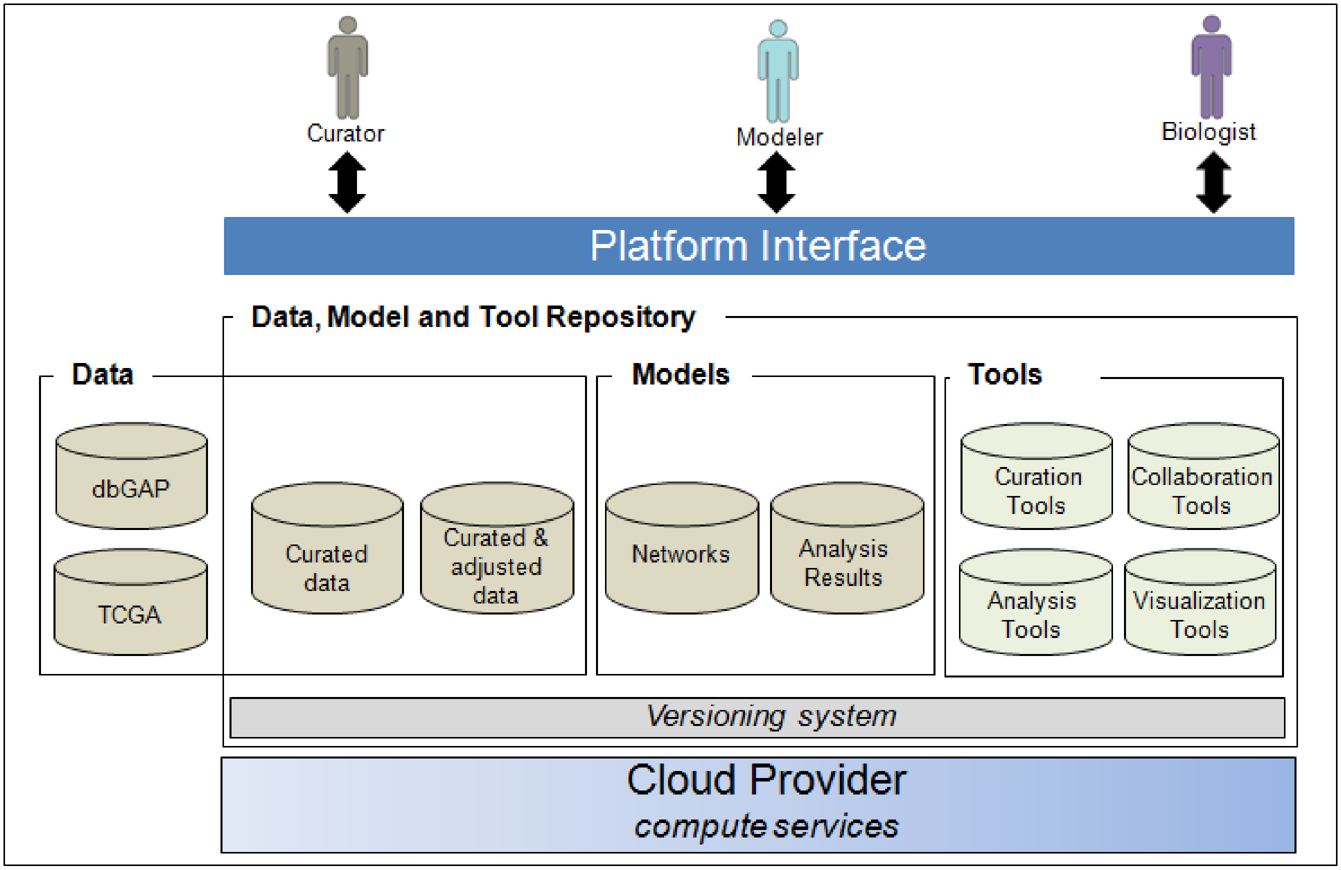

Figure 1: Sage Bionetworks Platform Architecture: The Platform Interface uses a set of web services to provide access to the Data Repository: a federated collection of curated, adjusted, and analyzed datasets, network models, and code. The Platform may also reference restricted data stored in external databases such as dbGAP or TCGA. In these cases, rather than providing data the Platform will provide a link to the data source and code necessary to process and curate the data into a standard format; the user can then apply this code to the data once they have received the necessary permissions through the established application processes from these external databases. All resources managed by the Platform can be referenced as objects via a URL following linked data principles. This approach allows storage of data and metadata using persistence mechanisms appropriate for each data modality, while abstracting multiple clients away from the details of how data and services are obtained. Integration with ontology services and support for a rich query language would occur on the Platform back-end, allowing multiple clients (e.g., R and the web client) to run similar queries across hosted data. Versioning of data, workflows, and tools allows the details of how individual network models are generated to be documented and hence reproduced. Storage of the data repository and services in the cloud allows for scalability, access, and the potential to use high performance compute facilities directly from the platform.

network models, and code constitutes a repository that scientists may access and reuse in new projects.

A set of applications will support various types of users, data curators and bioinformaticians to biologists, similar to open source development environments. A web portal under development will allow researchers to search and navigate through content relevant to their research interests and form projects with existing or new colleagues. General-purpose tools like wikis, user forums, and issue trackers can easily be adopted from other domains to support scientific research teams. More focused analytical work and scientific visualizations will be supported by integrating existing tools with platform services to provide access to hosted data in an environment already known and comfortable to analysts. For example, we have begun the development of an R package to allow platform-hosted data to be accessed in this open-source environment, and link to the wealth of existing analysis methodology available via resources like Bioconductor.

Access to shared data and functionality across applications are provided by a set of web services that provide a central location for a variety of features including annotation, indexing, history tracking, versioning, authentication/authorization, and data persistence. We have designed an application programming interface (API) that allows for structured queries including filtering, sorting and paging, across the metadata of all datasets, network models, and tools registered by the Platform. The API, inspired by Facebook's Graph API and Query Language and API[27], consists of both a set of simple calls to access individual datasets, plus a richer query against the metadata store. Structured query services can be semantically enhanced by having the Platform services delegate to external services, such as NCBO[28], to expand query terms prior to interrogating the internal persistence layer, e.g., to find datasets where the tissue is labeled as "Cerebrum" or "Cerebellum". A service-oriented architecture also has the advantage of helping to keep analysis code near the data, which is becoming ever more problematic to move as data volumes are growing faster than network bandwidth. Due to the scope of the data and analysis demands, the platform will be built to take advantage of the maturing set of cloud compute technologies that can provide scientists super-compute power on demand without the up-front capital costs of building and managing a private cluster.

Data and Network Repository: The data content underlying the Platform is stored in a centralized repository of curated data and network models (http://www.sagebase.org/commons/repository.php and Figure 2). The process of curation of acquired data involves both data integrity checks and transformation of the dataset into a standard format as described below. These curated datasets represent an essential building block for shared, versioned analyses. The platform will also provide a second "analysis-ready" version of each dataset that has been through a Quality Control (QC) process in preparation for modeling in conjunction with tools, source code and detailed documentation that link the adjusted dataset to the underlying curated dataset (Figure 2). We recognize that normalization and data adjustment processes may differ depending on analytical goals and so this effort will provide a mechanism for end users to create and store different versions of each dataset that include code written for alternative adjustment strategies. We welcome input from the community and contributions of raw, curated and/or adjusted datasets in conjunction with the relevant annotation and/or analysis tools.

Initial efforts to populate the Platform have focused on the collection and curation of multi-layered genomic datasets that are essential for the development of predictive computational models. Data are collected from diverse sources (Figure 2); from studies in public repositories as well as directly from investigators employed in either academic or commercial efforts. Although this project is not aimed at recreating the extensively populated genomic databases that already exist[29-32], we are working to improve the usability of publicly available datasets that are best suited for model building. Much of the data currently available through public repositories lacks critical information required for effective use; datasets may lack essential metadata or annotations, unique participant identifiers that map individuals across data-types, or adequate descriptions of data processing. In addition, data are not typically shared in a standard format, requiring significant effort for analysts to curate them. Furthermore, model building is most effective when using data containing multiple-layers of genomic and phenotypic information (e.g., DNA variation, expression profiles, and clinical traits) but these different types of data are often deposited across multiple repositories. For studies in non-human model systems, no repository exists.

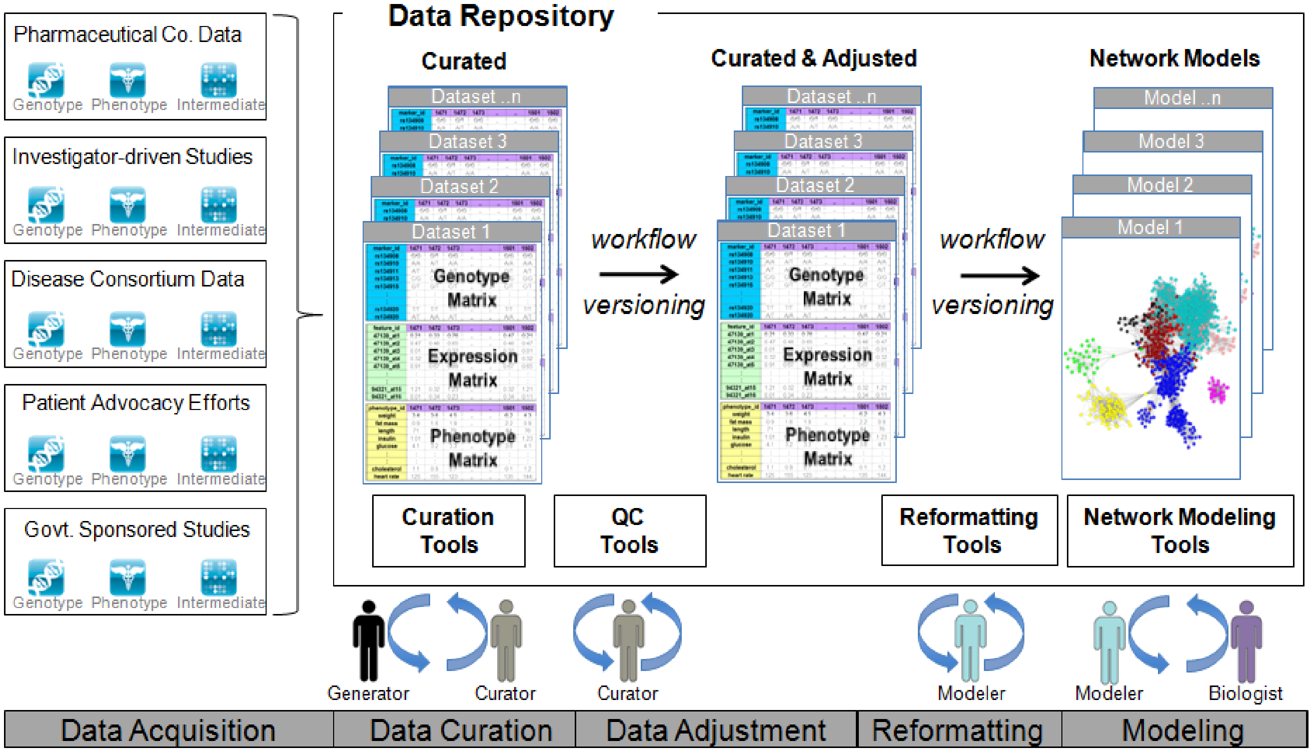

Figure 2: Process of Data Acquisition, Curation, Adjustment, and Modeling. Data flows into the repository from a number of different sources with examples listed. Individual datasets typically contain different data-types and are submitted in various formats. Curation involves reformatting into a common tab-delimited text matrix format. This curated standard format is available for download as well as allowing the development of workflows for common manipulations (e.g., adjustments for technical co-variates 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 important feature being that the version of the dataset that is used for an analysis as well as the code and workflows are stored. Allowing different types of users to interact with the data at different points has advantages. For example, providing tools to enable curation of a dataset into a standard format provides the user with the benefit of easy curation and opens up tools for downsteam QC and analysis. For the repository, the potential benefit in providing these tools may be in encouraging broader data sharing.

Datasets in the Platform repository are generally derived from human or mammalian studies containing more than 50 subjects: mammalian models tend to more accurately reflect the complex interactions observed in human disease and modeling techniques require a minimum number of individuals in order to build accurate models. Select genomic datasets that contain phenotypic traits in conjunction with either DNA variation or molecular trait data can be used within a subset of modeling techniques and may be included. Currently, molecular traits and genotype data are most commonly available as genome-wide gene expression profiles and SNP profiles respectively, but in principle any comprehensive measures of molecular phenotypes can be considered as they become available, including proteomic, metabolomic, whole genome and transcriptome sequencing, miRNA, and other types of molecular data. We encourage investigators to directly contact Sage Bionetworks regarding datasets or models to be shared (repdata@sagebase.org).

The curation of datasets involves a series of integrity checks and transformations into a standard format to maximize accessibility and utility. Specifically, the steps are: (1) collecting and inventorying the data; (2) verifying sample matching across data-types, including checking for inconsistencies in the data based on inferred gender and repeated samples; and (3) formatting of the data files to standard formats. To maximize flexibility, curated data is released in a basic tab-delimited format with each data-type stored in rectangular tab-delimited files, where individuals populate the columns and the molecular (SNP, gene expression reporter, etc) or clinical phenotypes populate the rows (Figure 3). Each data-type has a companion annotation file, which is also in tab-delimited format, where each row describes the molecular (probe, marker) or clinical (trait) phenotypes present in the data file. Sample annotations, such as batch information, are stored in the phenotype file along with the traits measured on each individual. While this manner of data storage requires some manipulation to conform to each of the commonly used data formats for downstream analysis (e.g., Bioconductor, Matlab, Plink), the standard and flexible nature allows tools to be readily written to perform that manipulation. These tools will be distributed through the Platform, as well as through other standard tool-sharing mechanisms. Standardizing formats and annotations will facilitate data discovery and re-use. Ultimately we aim to create a set of semantically rich annotations for curated data and validate these using ontology services (e.g., as supplied by the NCBO)[28,33] to ensure consistency of terminology across datasets. For clinical data, it will be necessary to also map clinical variables and covariates to ontologies in a similar fashion to facilitate meta-analysis across similar clinical studies. Curation also allows processes to be employed that aid in the protection of patient privacy[34]. A detailed description of the curation process and the resultant curated data packet including the content and format are available at the Sage Bionetworks website: http://sagebase.org/downloads/SageBio_DataSpecification_MAR2011.pdf.

The ultimate goal of this curation effort is to provide a mechanism for the collaborative generation, modification and improvement of predictive computational network models of disease. Networks are a convenient way to represent complex molecular interactions and to provide a framework for predicting causal relationships between molecular entities and clinical outcomes. However, the current standard of publishing modeling methods as general descriptions in manuscripts does not provide sufficient detail for a network to be accurately reproduced. Network models can be decomposed into a collection of triplets that define the relationship (or edge) between two nodes

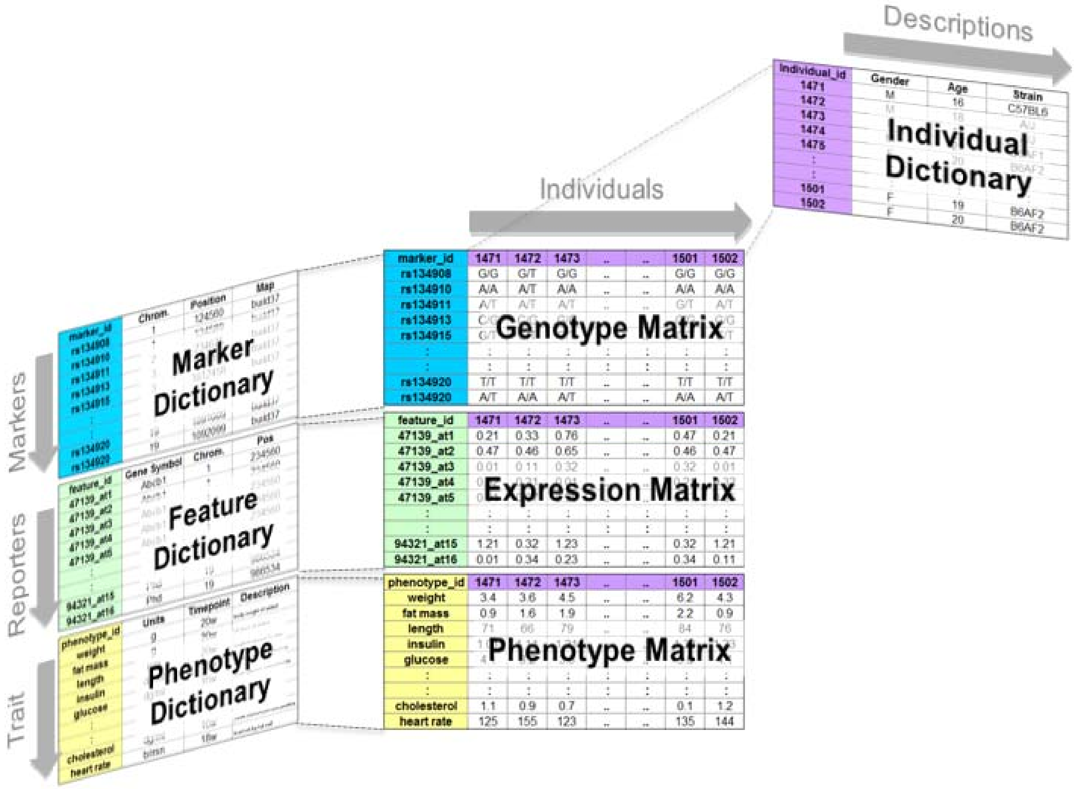

Figure 3: Graphical Representation of a Curated Data Packet. All data (matrix files) and annotations (dictionary files) are contained in tab-delimited text files. Each data-type has a matrix file containing the data-type identifier on the rows and the individuals on the columns. Associated with each matrix file is a "dictionary" file that provides annotations for that particular data-type. In addition, there is an individual dictionary file that describes the individuals in the study. A free text metadata file that describes the study design, data generation technologies, data contributor, and terms of data use accompanies each data packet. The example shown is a "standard" Affymetrix package with gene expression data, genotype data, and phenotype data; note in the Affymetrix technology individual sequences designed to hybridize to specific RNA species are termed features. More complex datasets may contain additional data-types and accompanying dictionary files.

where a node represents either a molecular entity such as a gene or protein or a phenotype such as biochemical measures or clinical outcomes (Figure 4).

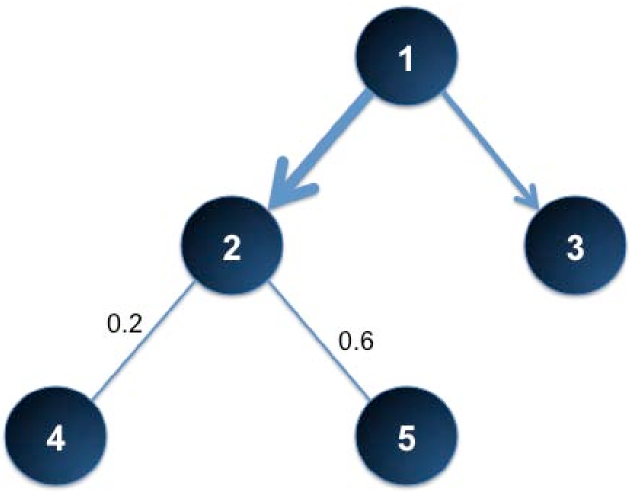

Figure 4: Example of a Network Model. A graphical network model describes the relationships between molecular or phenotypic entities. Typically this is modeled in the form of nodes representing genes, proteins, or metabolites, and edges or connections, describing the relationships between them. Relationships can be the correlation of continuous variables across a population, physical protein interactions, or other such dependencies. The graphical structure can convey additional information such as directionality, strength of a connection, and/or the confidence in a connection. In the example the direction of information flow through the network, derived from causal inference, is denoted by arrows, such that node 1 drives the behavior of nodes 2 and 3. Note the lack of edge between 2 and 3 indicates the effects of 1 on 2 and of 1 on 3 are independent of each other. The lack of arrows between 2 and 4 and 2 and 5 indicates no directionality due to uncertainty of the causal relationships. The strength of connections can be shown in terms of the width of edges (compare 1 to 2 versus 1 to 3) and the confidence in the existence of an edge by numeric terms (compare 2 to 4 versus 2 to 5).

This information can be conveniently shared through text files. This model format, as with the curated data format, has been selected because it is a readily useable, universally exchangeable format and tools that transform models from this format into other common formats can be developed as necessary. Distribution through the Platform of models linked to underlying data, detailed and versioned code, and analytical workflows, will ensure that models are built in a reproducible manner. This transparency is designed to encourage collaboration on the decisions made within the modeling process, to ensure that the quality of network models can be meaningfully assessed, and to provide a forum for the development of modeling standards and the sharing of modeling techniques. Interactive feedback will further drive this process and direct biological researchers to high quality models from which they can inform their own research efforts. Initially we will ask model generators to provide the source code used to generate each model but ultimately the Platform will be able to incorporate standardized modeling tools and track the methods used by the modeler. In addition, we recognize that other analyses and representations of the data beyond networks are also useful and we anticipate that the Platform will also support these.

Tool Repository: A key feature of the platform will be the tools developed both by Sage Bionetworks and by the broader scientific community for the manipulation and analysis of data (Figure 2). As an example, one of the significant bottlenecks in analysis of large-scale genomic data is the time and effort required to curate and QC the data. As an initial effort we aim to provide a standardized system for data curation with specific software tools to facilitate both the preparation of the basic curated data files and also their conversion to other data formats for downstream analysis. We envisage that this alone will be of significant value to the community; curation of data is often a time consuming and even daunting task for the less experienced investigator. Further, it has the potential to both incentivize data sharing and provide a mechanism for scaling the curation process by involving a broader group in the process. Additionally, tools for data adjustment and reformatting, as well as for model building, will be developed. These software tools (http://www.sagebase.org/research/tools.php) will be open source in the sense that the source code will be freely available under a license that not only allows, but also encourages, the sharing of the code and any enhancements made to it.

Data Sharing and Reproducible Science

In order to provide a community for collaboration, the data, tools and models shared through the Platform must be made broadly available with the fewest possible restrictions. It is essential that this resource be as accessible to a high school student developing a science project as to an academic investigator that specializes in modeling methodologies. For this reason, access to the Platform will be available to any registered user for research or training purposes. Users who want to download data will be authenticated via a valid name and e-mail address and will be encouraged to provide a short paragraph outlining analytical goals the first time they request access to a dataset. The identity of researchers and the analytical plans for each dataset will be made available to the community but will not be used to determine eligibility for data access.

Data sharing requires a willingness of data generators to provide access to data –something that is becoming more appreciated to maximize research impact and accelerate the rate of scientific discovery[18-22,35]. Although the benefits of sharing scientific data are widely acknowledged, the implementation of this process has lagged due to multiple technical and cultural challenges. Responsibilities and funding for the organization, storage, and curation efforts, a prerequisite for widespread data sharing, remain unclear. For this reason, data are often shared in a manner that is difficult to interpret or that lacks essential elements required for the data to be reused. One major barrier to data sharing is the perception that data analysts get a broader benefit from sharing of data than do data generators or curators. A truly open policy towards the rapid and broad sharing of data will necessitate novel mechanisms that properly acknowledge the efforts that go into data generation and curation. This will include formally linking these efforts to the datasets, journal citations, and recognition of data generation and sharing efforts within professional merit systems including qualifications for tenure advancement. Some of the solutions to broader data sharing are already being driven by publishers and funders in policies requiring data deposition prior to manuscript publication and/or the completion of a funding cycle[36-38]. Other initiatives such as mechanisms for attributions beyond standard print publications are likely to be important[39].

Sharing of human genomic data presents additional challenges. Common concerns include maintenance of the privacy of human participants and minimization of the risk that misuse of the data could lead to stigmatization or discrimination in insurance, employment or other situations. Participant health information is protected both by federal laws[40,41] and by state laws that can impose additional restrictions. In accordance with these rules all data available through the Platform will be stripped of identifiable information[41]. A number of efforts to change the patient consent process and allow patients to control how and with whom their data are shared are being discussed as a potential solution to guarantee an individual's right to selective disclosure of their data[42]. Another approach is to engage the communities of users and contributors to act as stewards of the data. The notion of data stewardship implies mutual investments and shared responsibility by everyone to protect the interests of study participants as well as the interests of investigators. In this new paradigm, shared expectations and trust among community participants are essential[43]. A resource like the Sage Bionetworks Platform should be used to increase knowledge as a benefit to the community within the context of respect for the values and intentions of study participants and in a manner that limits risk for misuse leading to population discrimination and/or marginalization. To this end, Sage Bionetworks will require that users of the Platform agree to certain Terms of Use and ethical behaviors such as the promise not to identify human participants and the agreement not to use the data in a way that would enable genetic discrimination (see: http://sagebase.org/downloads/SageBio_TermsOfUse_MAR2011.pdf). Data submitters may indicate additional terms and use restrictions to certain datasets in accordance with participant's informed consent directives, or on datasets whose full disclosure could carry risks for individual's privacy or group stigmatization. These terms or restrictions may vary by study to provide the appropriate protection of participants in each trial, i.e., limitation to certain fields of study, location of research, and analysis types. Some genomic repositories and institutions (e.g., dbGAP, ICGC) have developed their own data access policies and procedures to prevent the unrestricted dissemination and potential misuse of human genomic data, and where appropriate we will direct users to apply for data access through the appropriate established mechanisms selected by the data contributor (e.g., dbGaP or institution websites). In these cases only the software code and network models derived from the data will be shared directly through the Platform with a reference to the underlying dataset. Sage Bionetworks does not currently host restricted datasets but in the eventuality where restricted datasets are stored in the Platform, data users will be expected to fully comply with the limitations and restrictions set by the data submitter. Sage Bionetworks will not arbitrarily impose restrictions to data use. The consequences of violating these rules will be the denial of continued access to the Sage Bionetworks website

With greater openness comes the need for greater accountability and vigilance through community engagement. We propose to use a public forum to promote ethical behavior and prevent misuse of the data. The Platform will provide a way to give feedback and log in concerns so that issues, whether logistic, scientific, ethical or regulatory can be brought to the attention of the community and be rectified promptly. This model promotes a different culture and requires new governance guidelines. We highlight how Sage Bionetworks proposes to address these issues and the challenges associated with broad sharing of genomic data at the Sage Bionetworks website (see: http://sagebase.org/downloads/SageBio_Governance_MAR2011.pdf)

Future Directions

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

We believe this effort distinguishes itself from the data warehouse-based efforts at NCBI, EBI, and elsewhere[29-32] in going beyond the straight repository model to incorporate the concepts of community-based analysis and modeling. Other efforts to build model repositories, such as EBI's BioModel[44], have focused on deductive methodologies and defining standards for representing mechanistic models of biochemical systems. While we take inspiration from these sorts of resources, our focus on data-driven statistical models has led us to conceive of the platform as integrating both data and models in a single resource. Knowledge-based systems built on the current literature or data repositories, such as Ingenuity, Nextbio, or Gene Atlas also offer some of the tools but are focused on helping users mine existing knowledge using predefined analyses, rather than letting scientists explore novel mathematical approaches. We envision integrating results generated by platform scientists with a variety of these sorts of resources. Other efforts contain elements of the workflow and reproducibility themes central to the Platform, notably Gene Pattern[45], Taverna[46], or Galaxy[47] and these sorts of workflow systems and tool collections could be linked to the platform. Ultimately we believe that the combination of a repository of curated and readily usable data, along with tools for data manipulation and analysis, a workflow versioning system and forum for collaborative modeling provides significant attractive features for users of the Platform. The open architecture of the platform should also make it possible to link to tools developed by others and indeed the presence of a repository of standardized and curated data will make direct comparisons of these approaches more tractable[48-52].

As is natural for a project of this scope all the answers are not available at the start. It is possible that unforeseen modeling techniques will be developed that make obsolete the current state of the art methodologies. Indeed, this is our hope: that by opening up the data and current approaches more people will have ready access to them and out of this will come better models. Thus, the focus should be on enabling rather than competing, so that patients are the winners through access to better drugs. Perhaps the hardest challenges ahead do not relate to technical challenges but rather to sociological ones related to how we collaborate. Undoubtedly the incentives for broad data sharing will not work for everyone, but if we recognize the importance of this issue it can be solved over time as a community. Initially the drivers will be from the few who recognize the importance of this concept, backed up by policies from publishers and funders, but ultimately the drive for change will come from striking examples of what can be done with broad data access and crowdsourced analysis.

In conclusion, we believe that the time is right to combine the power of social networking, in terms of sharing and working as distributed teams, with the needs and opportunities engendered by the genomics revolution and the desire to translate public and private investment into demonstrable human benefit. Indeed, we believe that the scale and complexity of the problem ahead of us dictates that we work together in transparent and reproducible ways, and that we share and reuse data, tools, and models.

References

1. The human genome at ten. Nature 464, 649-50 (2010).
2. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates. Nature Rev. Drug Discov. 3, 711-715 (2004).
3. Tony Hey, Stewart Tansley, and K.T. The Fourth Paradigm: Data Intensive Scientific Discovery. (2009) at http://research.microsoft.com/en-us/collaboration/fourthparadigm/
4. Tegnér, J.N. et al. Computational disease modeling - fact or fiction? BMC systems biology 3, 56 (2009).
5. Barabási, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5, 101-13 (2004).
6. Bystrykh, L. et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using "genetical genomics". Nature genetics 37, 225-32 (2005).
7. Carro, M.S. et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463, 318-25 (2010).
8. Chen, Y. et al. Variations in DNA elucidate molecular networks that cause disease. Nature 452, 429-35 (2008).
9. Emilsson, V. et al. Genetics of gene expression and its effect on disease. Nature 452, 423-8 (2008).
10. Ghazalpour, A. et al. Genomic analysis of metabolic pathway gene expression in mice. Genome biology 6, R59 (2005).
11. Bandyopadhyay, S. et al. Rewiring of genetic networks in response to DNA damage. Science (New York, N.Y.) 330, 1385-9 (2010).
12. Schadt, E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nature genetics 37, 710-7 (2005).
13. Zhu, J. et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature genetics 40, 854-61 (2008).
14. Schadt, E.E., Friend, S.H. & Shaywitz, D.A. A network view of disease and compound screening. Nature reviews. Drug discovery 8, 286-95 (2009).
15. Friend, S.H. The need for precompetitive integrative bionetwork disease model building. Clinical pharmacology and therapeutics 87, 536-9 (2010).
16. The National Center for Atmospheric Research at <www.ncar.ucar.edu>
17. Netflix Prize. (2009) at <http://www.netflixprize.com/>
18. Hrynaszkiewicz, I. The need and drive for open data in biomedical publishing. Serials: The Journal for the Serials Community 24, 31-37 (2011).
19. Friend, S.H. Something in common. Science translational medicine 2, 40ed6 (2010).
20. Schofield, P.N. et al. Post-publication sharing of data and tools. Nature 461, 171-3 (2009).
21. Guttmacher, A.E., Nabel, E.G. & Collins, F.S. Why data-sharing policies matter. Proceedings of the National Academy of Sciences of the United States of America 106, 16894 (2009).
22. Field, D. et al. Megascience. 'Omics data sharing. Science (New York, N.Y.) 326, 234-6 (2009).
23. Mesirov, J.P. Computer science. Accessible reproducible research. Science (New York, N.Y.) 327, 415-6 (2010).
24. Gentleman, R. Reproducible research: a bioinformatics case study. Statistical applications in genetics and molecular biology 4, Article2 (2005).
25. Bukheit, J. & Donoho, D. Wavelab and reproducible research. Wavelets and Statistics 55-81 (1995).
26. Sage Congress. (2010) at <www.sagecongress.org>
27. Facebook Query Language at <http://developers.facebook.com/docs/reference/fql/>
28. The National Center for Biomedical Ontology at <http://www.bioontology.org/>
29. NCBI Database on Genotypes and Phenotypes at <www.ncbi.nlm.nih.gov/gap>
30. EBI Array Express at <http://www.ebi.ac.uk/arrayexpress/>
31. Pharmacogenomics Knowledge Base at <http://www.pharmgkb.org/>
32. NCBI Gene Expression Omnibus at <http://www.ncbi.nlm.nih.gov/geo/>
33. Jonquet, C., Musen, M.A. & Shah, N.H. Building a biomedical ontology recommender web service. Journal of biomedical semantics 1 Suppl 1, S1 (2010).
34. Hrynaszkiewicz, I. et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ 340, c181-c181 (2010).
35. Challenges and Opportunities. Science 331, 692-693 (2011).
36. Data's shameful neglect. Nature 461, 145 (2009).
37. Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. (2003) at <http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf>
38. NIH Data Sharing Policy and Implementation Guidance (2003) at <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm>
39. Giardine, B. et al. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nature genetics 43, 295-301 (2011).
40. Genetic Information Nondiscrimination Act. (2008) at <http://www.genome.gov/24519851>
41. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy and Security Rules. (1996) at <http://www.hhs.gov/ocr/privacy/index.html>
42. Shelton, R.H. Electronic consent channels: preserving patient privacy without handcuffing researchers. Science translational medicine 3, 69cm4 (2011).
43. Anderson, N. & Edwards, K. Building a chain of trust. Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies - GTIP '10 15-20 (ACM Press: New York, New York, USA, 2010).doi:10.1145/1920320.1920323
44. Biomodels Database at <http://www.ebi.ac.uk/biomodels-main/>
45. Gene Pattern at <http://www.broadinstitute.org/cancer/software/genepattern/>
46. Taverna Workflow Management System at <http:>
47. Galaxy at <http://main.g2.bx.psu.edu/>
48. Bioconductor at <http://www.bioconductor.org/>
49. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102, 15545-50 (2005).
50. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics 34, 267-73 (2003).
51. Vaske, C.J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics (Oxford, England) 26, i237-45 (2010).
52. Tarca, A.L. et al. A novel signaling pathway impact analysis. Bioinformatics (Oxford, England) 25, 75-82 (2009).