# 2. Specific Aims

In the IT industry, a wealth of new product launches is fueled every year by Moore’s Law increases in compute power. In the pharmaceutical industry, however, an exponential growth of biological “omics” data has correlated with a decline in the number of NMEs approved by the FDA. Despite increasing expenditure and research discoveries, the global biomedical research enterprise is failing to generate innovative products that solve unmet medical needs or replace the revenues lost as existing medications go off patent.[[1]](#endnote-1) Sage Bionetworks believes that a fundamental reason biological research productivity does not scale with biological data generation is that the analysis and interpretation of genomic data remains largely an isolated, individual activity. Pharmaceutical R&D pipelines, and even many pre-commercial research programs, consist of a series of handoffs among isolated scientists with different areas of expertise. The ability to access, understand, and reuse data, analysis methods, and models at each step is a rate-limiting step in research progress and in developing new cures and treatments for human disease. Precompetitive collaboration is essential to the rapid translation of biomedical discoveries and requires a shift from the traditional and largely unsuccessful single lab, single-company, and single-therapy R&D paradigm.

Sage Bionetworks’ mission is to catalyze a transition to large scale, cooperative and distributed data analysis in human health sciences. For this to happen, it is critical that: 1) human health data become accessible and reusable by people other than the original data generators so that multiple approaches to data interpretation can occur in parallel 2) analytical methodologies be fully reproducible and transparent so that results can be vetted and existing analysis techniques quickly applied to new application areas, and 3) models of biological systems and networks be opened to a variety of users such that theoretical predictions can be rapidly validated experimentally and improve standards of care for patients. Sage Bionetworks is actively engaged in developing solutions to each of these issues with its academic and pharmaceutical collaborators. To support these efforts, the PI of this grant is leading the development of the Sage Bionetworks Platform, which will provide informatics support for both Sage’s research initiatives and the broader scientific community. This proposal would fund a Sage Bionetworks-based Driving Biological Project (DBP) that would broadly apply National Center for Biomedical Ontology (NCBO) technology to help Sage achieve its mission.

**Aim 1: Embed NCBO technology throughout the Sage Platform to facilitate curation, discovery, analysis, and reuse of Sage-hosted global coherent data sets and network models.** The Sage Platform will support the reusability of information facilitated by ontology-based services and applications directed at scientific researchers and data curators.

**Aim 2: Use enrichment analysis to dissect relevant substructures in biological networks.** Understanding how network structures relate to disease and response to treatments is a core area of research at Sage Bionetworks. We expect Sage’s need to classify regions of networks or gene signatures by various ontologies to use and drive the development of gene set enrichment analysis tools by NCBO.

# 3. Research Strategy

# Background and Significance

Each year the US pharmaceutical industry sets new records for R&D expenditure, rising from approximately $10 billion annually in 1990 to an estimated $65 billion in 2009. Sadly, this increase in investment has not been matched by increases in new drugs brought to market as evidenced by the fact that the overall number of new drugs registered annually is similar now to 20 years ago. This is attributed to the high attrition rate of drugs in clinical development with an increasing failure of drugs at Phase II proof of concept1,[[2]](#endnote-2). Indeed, the overall success rate from first-in-man to an approved drug stands around 10% with the major causes of failure being lack of efficacy and toxicity. These failures drive up the cost of developing a drug, now estimated at a staggering $1 billion[[3]](#endnote-3), and contribute to the spiraling health costs and lack of progress for the patient.

Recent high profile failures are exemplars of this problem. Because high-density lipoprotein cholesterol (HDL-C) levels are inversely related to cardiovascular disease (CVD), it has been assumed that raising HDL-C levels would be beneficial. However, the recent failure of a cholesteryl ester transfer protein (CETP) inhibitor (torcetrapib) developed by Pfizer to decrease CVD raises questions about this whole strategy[[4]](#endnote-4). Despite spending ~ $800 million it is still unclear whether the strategy of targeting CETP is flawed or alternatively if torcetrapib has drug-specific, off target, non-CETP dependent effects. It is a damning indictment of the clinical research enterprise that this vast sum of money could have been spent without reaching a conclusion about even the validity of the drug target. Why and how did this happen? We suggest that this occurred because the clinical data and study were structured and executed with only one goal in mind: to determine if one particular pharmaceutical company could or could not market one particular drug for one particular condition.

# Prior Work

Sage Bionetworks is actively pursuing the acquisition, curation, statistical quality control, and hosting of human and mouse global coherent datasets for use by Sage Bionetworks researchers, collaborators, and the broader research community. The datasets contain clinical phenotypes and genomic data, and an intermediate layer. Typically studies currently contain genome-wide genetic variation data (typically SNP and/or CNV) and/or expression profiling (typically mRNA microarray) but other data modalities could become prevalent as next-generation technologies mature. Current Sage Bionetworks efforts are focused on public hosting all of these datasets and the derived disease models, with prioritization of datasets containing multiple types of genomic data. Much of this work is supported by grants from the National Cancer Institute Integrative Cancer Biology Program and the Washington Life Science Discovery Fund.

Data generation efforts within the academic community are increasing exponentially; nevertheless the pharmaceutical industry remains an even larger, untapped and powerful resource for large scale clinical and molecular datasets from their development and trial activities. Through Sage Bionetwork’s legacy as a former research group of Merck & Co., Inc., and Sage’s President’s previous position as global head of oncology for Merck and extensive connections in other leading firms, Sage Bionetworks is uniquely positioned to access data from the pharmaceutical industry that would previously never been released. Indeed, Sage Bionetworks is engaged in a public-private partnership project to provide public access to genomic datasets generated within the comparator arm of industry-sponsored clinical trials and to combine them with public datasets generated by academic consortiums to advance understanding of both disease states and treatment regimens. These will be represented by two major types of datasets: (1) datasets that pair clinical traits with gene expression profiles, typically profiling tumor tissue samples from oncology trials, and (2) datasets that pair clinical traits with genetic variation data, typically genome-wide SNP panels from trials in fields other than oncology.

An example of the richness of these two types of data is an oncology dataset to which we have authorized access from GlaxoSmithKline that includes genome-wide expression profiling and clinical data collected from comparator arm participants in a Phase 3, randomized, open label trial comparing 3rd line treatments (capecitabine ± lapatinib) for HER2-positive advanced breast cancer (clinicaltrials.gov #: NCT00078572)[[5]](#endnote-5),[[6]](#endnote-6). Capecitabine is a pro-drug of 5-fluoruracil that is marketed by Genentech, a member of the Roche group, and was administered to patients using standard methods (2500 mg/m2 on days 1-14 of a 21 day cycle). Genome-wide expression profiles performed using the Affymetrix HG-U133 2.0 platform are available from 45 tumor samples collected at the time of trial randomization. A rich collection of phenotypic traits are also available for these participants including demographics, prior disease and treatment status, lab measurements collected during treatment, disease characteristics including classification of tumor using RECIST criteria, complete metastatic characterization, adverse events, quality of life descriptions and survival characteristics. Molecular phenotypes include HER2, estrogen receptor and progesterone receptor status of tumors.

We expect the release of these sorts of unique, integrative, high value datasets into the public domain can seed a variety of analytical approaches to drive new treatments based on better understanding of disease states and the biological effects of existing drugs. Indeed, it is the potential for such increased productivity that is motivating Sage Bionetworks’ pharmaceutical partners into beginning to release data that they have previously been jealously guarding.

Over the past eight years, the Rosetta/Sage group has developed an integrated approach to exploring the molecular mechanisms that drive disease patho-physiology. It should of course be noted that this is one of several examples of using these computational methods to analyze complex datasets. The Sage Bionetworks strategy has in particular focused on developing disease maps or networks that allow molecular phenotypes to be causally linked to disease outcomes and addresses many of the limitations of genetic association and linkage studies that simply link DNA variation to phenotypic measures without providing mechanistic insight. We have assembled human and mouse cohorts, in which tissues relevant to a diversity of human diseases such as obesity, cardiovascular disease, diabetes, atherosclerosis, chronic obstructive pulmonary disease, and asthma have been carefully collected in combination with physiological outcomes. By performing genome-wide genotyping of DNA collected from each individual in these cohorts and genome-wide profiling of RNA isolated from each tissue, the group has successfully integrated these data with a wide array of clinical and physiological phenotypes collected in the same cohorts and identified and validated both single genes and networks of disease. These data have been used to identify and validate a large number of genes for atherosclerosis, diabetes, and obesity related traits[[7]](#endnote-7),[[8]](#endnote-8),[[9]](#endnote-9),[[10]](#endnote-10),[[11]](#endnote-11),[[12]](#endnote-12),[[13]](#endnote-13),[[14]](#endnote-14) and leveraged to infer the causal genes affected by genetic loci identified in large-scale GWAS and provide a functional context within which to interpret those findings[[15]](#endnote-15).

There are several different network models that describe different aspects of biological systems. For example, co-expression networks provide global views of how biological systems are organized into different biological processes11,[[16]](#endnote-16),[[17]](#endnote-17) while probabilistic graphic networks, such as Bayesian networks, elucidate how genes are causally related to biological processes14,18,[[18]](#endnote-18). We have developed a logical analysis flow that leverages the power of both co-expression and Bayesian models. In the first step we construct complete co-expression maps of all the genes in a particular tissue across a population of interest. The individual groups of genes or modules that are identified by this method can then be annotated in two important ways: by reference to curated literature gene sets such as the Gene Ontology (NCBI) or KEGG pathways database; and by examining how predictive the expression behavior of the set of genes is for a particular clinical or phenotypic measurement using correlation of the eigengene(s) for the module with the phenotype values. This informs the researcher how the modules identified relate to known functions and more importantly partitions the modules allowing focus on those most correlated with phenotypic measures. In the second step of the method Bayesian modeling can be used to infer the regulatory structure of disease-correlated co-expression modules and the key drivers or hubs of the sub-networks identified. These key drivers represent key intervention points to modulate network activity and hence phenotypic outcomes as shown by the strong enrichment for these genes in positive hits from relevant siRNA screens (Zhang et al, in preparation). The publication of these findings in high impact journals such as Nature, Nature Genetics, PLoS Biology, and PLoS Genetics has firmly established this approach. Importantly this approach by credited by Merck executive management as supporting critical decision-making for a number of Merck developmental compounds.

Achieving the goal of data, analysis, and model reuse requires the development of a Sage Platform: a supporting informatics infrastructure to facilitate not just access to resources, but true re-usability. In this system, users interact with resources in the Sage platform via a number of mechanisms depending upon their interests and expertise. The system will provide scientists a means to easily search and navigate through content relevant to their research interests. The Sage Commons portal will be a “Web 2.0” environment for end user scientists to interact and share data, models, and analysis methods, both in the context of specific research projects, and broadly across otherwise disparate projects. Many other specialized scientific tools can be easily extended to load data and save results to the Sage platform, or to perform analysis by calling methods executed on a remote service. These more specialized analytical clients would support use cases in data curation and QC, as well as scientific analysis.



**Figure 1: Sage Platform Architecture and User Groups**

In the Sage Platform architecture (Figure 1); a set of REST-based web services provides access to the Sage Repository: a federated collection of curated, QC’d, and analyzed data, network models, and code. All resources managed by the Sage platform can be referenced as objects via a URL following linked data principles. This approach lets us store data and metadata using persistence mechanisms appropriate for each data modality, while abstracting our multiple clients away from the details of how data and services are obtained. We expect that integration with ontology services would occur on the Sage Platform back-end, as our services would delegate to NCBO services to access controlled vocabularies and ontologies, validate data, and semantically-enrich queries. Again, this approach is designed to keep business logic in one place as we have use cases for multiple clients (e.g. R and the web client) that need to run similar queries across Sage data.

The platform service layer provides a single place for a variety of general-purpose platform features:

* **Annotation** – A managed set of properties for all Sage resources that describe their structure and context.
* **Indexing** - Both structured and unstructured query mechanisms to find resources via indexes created by this layer.
* **History Tracking** – A recording of the history of who did what to produce a particular resource, resulting in a high level, uncurated, work flow for projects run on the platform.
* **Versioning** – Object level version history, with relationships between resources tracked at the version level
* **Authentication / Authorization** – Resource level control on user level access and guest level access that can evolve over time to reflect the changing nature of resource availability with project life cycle. This layer will leverage emerging standards (e.g. Open ID) to manage access to platform resources.

Additionally, the platform will require optimized storage and specific services for each of the types of resources hosted:

* **Data Repositories** provide access to structured data, stored either in flat-file, analysis ready binary object (e.g. R data file), or relational database format. Large-scale curated and QC’d data sets are likely to remain as either flat files or R binary data objects accessed through the platform as starting points of analysis. In some cases, portions of these data sets or analyzed results may be stored in a relational database to allow more advanced query and analysis of the data by scientific applications.
* **Model services** deal explicitly with managing biological network data. Semantic web technologies provide a natural framework for integrating network-centric data that may originate from a variety of sources. By storing network data as RDF and leveraging the query and inference capabilities of general-purpose triple stores, the platform will provide facilities for import and export of network data, searching networks for patterns of interest, and comparing and merging different networks. Exports to tools such as cytoscape will also be available.
* **Algorithms** will be stored in a code versioning system, providing standard software development tooling around algorithm development and release. Additionally, users or applications will have the ability to call algorithms, with appropriate job management and resource provisioning for large computation.

This service-oriented architecture has been designed leveraging ongoing dialogs with several groups such as Amazon, Microsoft, NextBio, the Institute for Systems Biology, and the Bioconductor group at the Fred Hutchinson Cancer Research Center. Each of these groups has significant expertise in various areas relating to platform development or bioinformatics, and we expect to build the platform in collaboration with some combination of these or other partners. Additionally, Sage Bionetworks has recruited a team of 5 professional software engineers with over 50 years total industry experience to design and build the platform, a subset of whom are named as resources on this grant. Funding is in place from the Washington Life Science Discovery Fund to support the development of the platform. Development is in progress; by the end of 2011 the essential components of the platform key to hosting the datasets as described in this proposal will have all been launched.

# Research Plan

**Aim 1: Embed NCBO technology throughout the Sage Platform to facilitate curation, discovery, analysis, and reuse of Sage-hosted global coherent data sets and network models.** At the core of many of the use cases the Sage Platform will support is the reusability of information facilitated by ontology-based services. We expect to roll out platform functionality incrementally, initially to Sage scientists, then to collaborators, and finally the broader scientific community. Short release cycles following agile development methodologies will be used to tailor the platform to the actual needs of real users. In this section, we describe how we will use NCBO technology to facilitate a variety of aspects of the use cases intended to be supported by the Sage Platform.

One area where the use of NCBO technology will come immediately into play is the curation of datasets hosted by the Sage Platform. In year one of the grant, we expect to focus on a curation use case in which Sage Bionetworks employees act as the primary curators of datasets to be hosted by the Sage Platform. Given the definition of a Sage global coherent dataset as a collection of heterogeneous layers of data on a set of common individuals, it becomes almost impossible for Sage to rely on any particular single data format as a curation standard as in single-datatype repositories like GEO or dbGaP. Rather, the strategy is to integrate with existing standards where applicable (e.g. CDISC for clinical data or MIAMI for microarray data), but to also make sure all annotations and descriptions of data are linked to appropriate ontologies to facilitate data discovery and re-use. For high-throughput data, it is sufficient to create a set of semantically rich annotations of the raw data, validated using NCBO ontology services, to ensure consistency of terminology across datasets. Annotations of datasets can then be stored and queried using a variety of approaches. We are currently evaluating the emerging class of schemaless data stores backing many large scale commercial websites (e.g. Google’s BigTable or HBase / Hadoop) as well as more conventional relational-database approaches for metadata storage. 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. Here, partnering with existing open-source clinical data management systems (e.g. I2B2) may be a viable option.

Given the breath of data hosted by the platform, we expect our curators to become actively involved in the selection and development of appropriate ontologies to catalogue the content. It will become necessary to allow Sage curators to provide detailed and precise feedback to the authors of existing ontologies on any areas where the ontologies do not seem to adequately represent how Sage datasets are to be annotated. Linking Sage curators into the ontology development process should provide driving use cases for the development of capabilities to support ontology life cycle management. This is an area of interest to Sage because we expect considerable evolution in curator’s requirements as we are still in the initial stages of the project, and expect to support new types of data as measurement technologies develop (for example, we are starting to see a transition from whole-genome SNP to whole-exome or whole-genome sequencing).

If the history of other public biological databases is any indication, Sage will quickly hit a stage where in-house curation efforts are insufficient to keep up with a rising flood of incoming data. Therefore, it is critical that any solutions developed for internal curation use cases in year 1 be constructed so that Sage can pivot towards a “crowd sourced” curation strategy by year 2. For a disparate and distributed group of end users to effective help categorize content will require the ontologies to be used to be well established, and the usability of the software tools to be much higher. We would also expect use cases involving expert review of new user’s submissions to become important to support.



Figure 2: Current wireframe mockups of Sage Platform have been used to validate software functionality with end users prior to engaging software engineering efforts. Interview with a variety of users have uncovered the need to support semantically-aware queries across platform-hosted datasets.

Discovery of datasets by end users is another key use case for the platform. Current wireframe mockups of Sage Platform show an interface where users can “shop” for data of interest; we expect the production web client to leverage the Google Web Toolkit and Google Visualization APIs to provide an easy-to-use interface for scientists to query for and discover datasets of interest. Widgets for auto-completion of terms or navigation of ontological hierarchies will be important to allow users to quickly narrow content to specific areas of interest. It is also crucial that queries leverage appropriate ontology hierarchies to capture all data of interest. A user looking for data from brain tissue should be able to find studies annotated with “cerebrum” and “cerebellum” as source tissues. Similar functionality will allow users to find network models or analysis routines of interest to their research.

Reusability of the query logic is achieved by separating the presentation layer of the web client from a REST-based service layer. We have designed a REST-API that allows for structured queries including filtering, sorting and paging, across the metadata of all datasets, network models, and tools registered by the Sage Platform. The API consists of both a set of simple CRUD calls to access individual datasets following conventional REST design patterns, plus a richer Query API also implemented as a HTTP-based service, but also taking a SQL-inspired attribute to specify a query against the metadata store. The design is inspired by Facebook’s Graph API and Query Language and API[[19]](#endnote-19). For example to find all data sets collected on brain tissue a client would issue a HTTP request of the form:

GET http://platform.sagebase.org/repo/v1/query?SELECT+\*+FROM+dataset+WHERE+'TissueType' = 'Brain'

It is important to note that although the query API looks like SQL, we don’t expect it to necessarily translate directly to a particular relational schema. Instead we are using a SQL-like language to express queries because SQL is commonly understood language and we expect it to be easy for external developers to use. Structured query services can be semantically enhanced by having the Sage Bionetworks services delegate to NCBO services (or, if necessary, a platform cache of NCBO data) to expand query terms prior to interrogating our own persistence layer.

Alternatively, Sage Bionetworks web client users might choose a free-text search over Sage content which could also leverage NCBO technologies to provide improved search results. In this use case, an end user would use the NCBO autocomplete widget to be guided to select terms that were defined in appropriate NCBO-hosted ontologies. We would also expect the Sage Platform to use the NCBO Annotator service to help index key concepts in free-text portions of our metadata against the same set of ontologies. By matching user queries and free text indices to the same set of terms, and leveraging synonyms and term hierarchies as defined in the ontologies, this approach should deliver increased relevance of search results.

The NCBO’s Resource Index is another mechanism by which potential Sage end users could be made aware of Sage-hosted data and networks. This mechanism is particularly interesting to us as this mechanism might bring new users to the Sage platform who otherwise might be unaware of its existence. We have already developed a REST-API for accessing Sage content with one target use case being the indexing of Sage-hosted content by a 3rd party. Since our own annotations of Sage-content will be validated against NCBO-hosted ontologies at create / edit time, building a spider that can index Sage content and populate the NCBO Resource Index should be straightforward.

**Aim 2: Use of enrichment analysis to dissect relevant substructures in biological networks.**

Gene set enrichment (GSE) analysis is an important technique for understanding broader trends in high-throughput gene expression data, such as pathway activity, cellular processes, or disease states. An underlying premise of GSE is that the concerted action of a set of genes is more closely aligned with molecular function and activity than any single gene, and is more likely to result in a stable and interpretable summary of the data. Scientists at Sage Bionetworks are frequent users of GSE methodologies, and have authored several GSE tools themselves.[[20]](#endnote-20),[[21]](#endnote-21)

A critical component for the successful utilization of GSE techniques is the gene sets themselves, derived from diverse sources such as expert curators, or machine learning methods. Among the most commonly used databases of gene sets are the gene ontology database and the pathway databases such as KEGG, Biocarta, and Reactome. In addition to these databases, the MSigDB hosted by the Broad Institute defines gene sets derived from perturbation experiments. While these gene set databases have been used in numerous studies, and have demonstrated their utility in many different biological contexts, they are by no means exhaustive and contain many biases such as over-representation of metabolic and cancer related pathways. By integrating with NCBO technology and its ontology framework, there is the potential to overcome some of these shortcomings. Using NCBO ontological framework, we would have access to a more diverse collection of gene sets, as well as the potential to derive new gene sets that are more relevant to different biological or clinical domains. A second advantage will be more precise filtering of gene sets to those that are relevant to the domain of study. This is significant methodologically, as filtering reduces the number of statistical tests and can help ameliorate the challenges imposed by multiple hypothesis correction.

Sage has several active studies and collaborations that will be utilizing GSE techniques. Among these is an aging project where we collect large human data sets and generate predictive models of aging. Gene set analysis will be required in this study to understand the underlying pathways up or down regulated in different human tissues, as well as any correlation these might have with disease phenotypes.

# Proposed Use of NCBO Technology

This research project will employ a variety of NCBO technologies to deliver functionality to our users as part of the public Sage Bionetworks Platform:

* NCBO type ahead and ontology query services to support free-text search of data, tools, and models hosted by the Sage Platform (Year 1)
* Ontology services to support tools for data curation, including collaborative development of new mechanisms for creating feedback loops between Sage data curators and ontology developers. (Year 2)
* Use of NCBO-hosted ontologies to structure data set annotations (Year 1) and clinical trial data (Year 2) to facilitate data discovery and meta-analysis
* Use of the annotator service for indexing free text descriptions of datasets and other Sage Platform resources (Year 1)
* Exposure of Sage content to the NCBO resource index (Year 1)
* Use of NCBO services to support GSEA by scientific research teams (Year 2)
* Scale testing NCBO technology in public cloud environment (Years 1&2, as part of each new platform feature area)

# Other Collaborative Considerations

Background of Sage Bionetworks: Sage Bionetworks is a new biomedical research organization formed in 2009. It is an independent corporate entity that leases offices at the Fred Hutchinson Cancer Research Center in Seattle, Washington. Sage Bionetworks is an IRS 501(c)(3) public charity incorporated as a non-profit in Washington State. Sage Bionetworks was formed as a strategic nonprofit research organization with a mission to coordinate and link academic and commercial biomedical researchers through a Commons that represents a new paradigm for genomics intellectual property, researcher cooperation, and contributor-evolved resources.

Sage Bionetworks works through partnerships with three primary activities: 1) Active collaborations with academic and commercial partners to apply advanced [integrative genomic analysis](http://www.sagebase.org/research/index.php) to genetic and clinical datasets; 2) training interdisciplinary researchers to acquire next generation bioinformatics skills (the focus of the Sage Bionetworks [National Cancer Institute Center for Cancer Systems Biology](http://www.sagebase.org/training/ccsb.php), and; 3) catalyzing and coordinating the development of a major new biological network and systems biology resource, the [Sage Commons](http://www.sagebase.org/commons/index.php).

Sage Bionetworks has been fortunate to develop a portfolio of funding including over $2m of philanthropic donations, over $13m of competitive federal and non-federal research grants, and over $4m of corporate partnerships that will assure continued growth and operation throughout the term of this proposal. It is worth noting that the legacy of core staff coming from a biotech division of a pharmaceutical corporation as well as the success in new commercial partnerships helps Sage Bionetworks maintain an outcomes and a customer focus even as its researchers work at the innovation forefront of computational biology. The active scientific projects at Sage put the platform development team in direct contact with a variety of researchers in both academic and commercial settings. This ready access to an active user base is a key ingredient in a successful software engineering project. Although no funds are requested to directly support any particular scientific project, the presence of Drs. Derry and Guinney as advisors to the software engineering team will ensure that the technology development supported by this proposal is directed towards having an impact on a variety of projects impacting human health.

Recently there has been an accelerating trend in the IT industry towards “cloud computing” environments in which large service providers (e.g. Amazon, Microsoft, and Google) provide on-demand access over the internet to shared pools of computing resources that can be provisioned, used, and released as needed. In addition to becoming an increasingly cost-effective strategy to provide compute and storage resources, cloud computing puts many basic IT management tasks (e.g. maintaining hardware, patching software, backing up data) into the hands of the cloud provider. As a new development effort with little existing legacy, the Sage Platform is aggressively leveraging and optimizing its architecture to take full advantage of these existing, and coming services, thus keeping Sage focused on tasks that require scientific expertise. Sage has developed a particularly strong IT partnership with Amazon, whose AWS team is conveniently headquartered less than a mile from Sage’s offices in the Fred Hutchinson Cancer Research Center. Additionally, Dr. Deflaux previously spent 7 years as a Principal Software Development Engineer for Amazon, where she helped design and scale Amazon’s IMDb, Mechanical Turk, and Search Inside the Book services. The close technical collaboration between Sage and Amazon will ensure that the Sage Platform is built for scale, and any scalability or other issues associated with using NCBO services within this cloud environment will come back to the NCBO as requirements and suggestions for design improvements.

It is also worth noting that the grant PI (Michael Kellen) and senior software engineer (John Hill) previously worked for Teranode Corporation, where Dr. Kellen led a software engineering team that developed Fuel. This product was a semantically-enhanced collaboration framework that leveraged ontologies to provide solutions for knowledge management in the pharmaceutical industry. In this role, both developed hands-on experience working with W3C’s semantic web standards including RDF, OWL, and SKOS, as well as several relevant public ontologies including the NCI Thesaurus, GO, and SNOMED.

**5. Bibliography and References**

1. Paul, SM et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* 9, 203-214 (March 2010) [↑](#endnote-ref-1)
2. Kola, I. The state of innovation in drug development. *Clin Pharmacol Ther* **83**, 227-230 (2008). [↑](#endnote-ref-2)
3. Adams, C.P. & Brantner, V.V. Spending on new drug development1. *Health Econ* **19**, 130-141 (2010). [↑](#endnote-ref-3)
4. Joy, T. & Hegele, R.A. The end of the road for CETP inhibitors after torcetrapib? *Curr Opin Cardiol* **24**, 364-371 (2009). [↑](#endnote-ref-4)
5. Cameron, D.*, et al.* Lapatinib plus capecitabine in women with HER-2-positive advanced breast cancer: final survival analysis of a phase III randomized trial. *Oncologist.* **15**, 924-934. Epub 2010 Aug 2024. (2010). [↑](#endnote-ref-5)
6. Geyer, C.E.*, et al.* Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N Engl J Med.* **355**, 2733-2743. (2006). [↑](#endnote-ref-6)
7. Chen, Y.*, et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429-435 (2008). [↑](#endnote-ref-7)
8. Emilsson, V.*, et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428 (2008). [↑](#endnote-ref-8)
9. Ghazalpour, A.*, et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2**, e130 (2006). [↑](#endnote-ref-9)
10. Zhu, J.*, et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* **40**, 854-861 (2008). [↑](#endnote-ref-10)
11. Schadt, E.E.*, et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003). [↑](#endnote-ref-11)
12. Schadt, E.E. Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr Opin Biotechnol* **16**, 647-654 (2005). [↑](#endnote-ref-12)
13. Yang, X.*, et al.* Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet* **41**, 415-423 (2009). [↑](#endnote-ref-13)
14. Zhu, J.*, et al.* Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. *PLoS Comput Biol* **3**, e69 (2007). [↑](#endnote-ref-14)
15. Schadt, E.E.*, et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008). [↑](#endnote-ref-15)
16. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005). [↑](#endnote-ref-16)
17. Emilsson, V.*, et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428 (2008). [↑](#endnote-ref-17)
18. Lum, P.Y.*, et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J Neurochem* **97 Suppl 1**, 50-62 (2006). [↑](#endnote-ref-18)
19. <http://developers.facebook.com/docs/reference/fql/> [↑](#endnote-ref-19)
20. Zhong, H., *et al*. Integrating Pathway Analysis and Genetics of Gene Expression for Genome-wide Association Studies. *American journal of human genetics* **86**, 581-591 (2010). [↑](#endnote-ref-20)
21. Edelman E., *et al*. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* **15**;22(14):e108-16 (2006).

**6. Protection of Human Subjects**

The research and collaborative activities proposed in this application involve the use of existing datasets and in all cases there are established and approved procedures in place to prevent the direct or indirect identification of subjects through the datasets or identifiers associated with the datasets. Data analysis and sharing procedures at Sage Bionetworks have been reviewed by the Western Institutional Review Board that concluded the activities are exempt. The proposal therefore is not considered human subjects research under 45 CFR 46.101(b)(4) which includes, “Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.” [↑](#endnote-ref-21)