NIH BD2K Centers of Excellence

Wow Stories

Feb 20th, 2017

TABLE OF CONTENTS

The BD2K Centers of Excellence appreciate this opportunity to share our contributions and achievements with the NIH Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI). Since its inception in October 2014, our program has made monumental progress in many aspects of data science and have achieved many of our major milestones. In doing so, we have developed tremendous trans-NIH momentum and synergy for furthering future innovations of greater value to science and medicine.

- These accomplishments are evident in our 546 publications and our release of 157 software platforms, computational pipelines, and a variety of computational resources (See Fig.1 below & **Appendix I**).
- These publications and products advance several key areas in data science. They collectively impact data annotation, data management, metadata standards, data integration, personalized real-world monitoring, predictive models, causal analytics, and precision medicine.
- In parallel to program-wide Consortium activities within the BD2K Centers of Excellence (See Fig. 2 below), Centers have collaborated on efforts with the NIH LINCS as well as the NIH Human Microbiome programs.
- In a survey conducted by the CCC, Center PIs affirmed that there will be direct applications of the tools to different disease areas, which support the mission of the following twenty ICs: NCI, NEI, NHGRI, NHLBI, NIA, NIAAA, NIAID, NIBIB, NCATS, NIAMS, NICHD, NIDA, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NIMHD, NINDS, and NLM (See **Appendix II**).

Highlighted in the one-page “Wow Stories” are accomplishments of Centers of Excellence, which exemplify the substantial progress of the thirteen centers. Moreover, we have identified opportunities for further advancement and have paved actionable steps for the next phase. We are committed to continuing our efforts to harmonize the tools in order to achieve interoperability and seamless integration of functionalities. We will continue to forge new advances in the specific applications of tools to propel our understanding of disease and biological pathways. We are confident that continued support of BD2K will powerfully and more fully harness the benefits of our products in a way that maximally impacts global science and health.

BD2K CoE	# Publication	# Software
BDDS	38	12
BDTG	56	31
CCD	23	5
CEDAR	6	1
CPCP	47	8
ENIGMA	154	6
HeartBD2K	55	37
KnowEng	34	7
LINCS-DCIC	38	23
LINCS-TG	1	2
MD2K	51	2
Mobilize	26	20
PIC-SURE	17	3
TOTAL	**546**	**157**

Fig 1. The number of publications and software tools that each center has created in furthering biomedical Big Data. Time of release is 2015-current (See **Appendix I**).

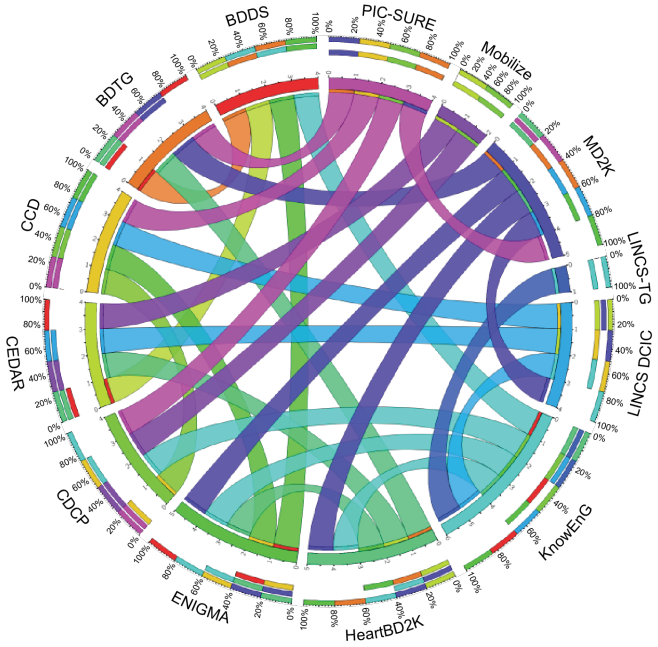


Fig 2. A collaboration map illustrating the partnership among the 13 Centers and their related work.

The Big Data for Discovery Science Center (BDDS) is a unique effort focused on the user experience with big data. How are big data organized, managed, and stored? How are big data processed and distributed either to local or remote computing resources or to colleagues or to geographically distributed archives? How can a user focused on biology interact with vast collections and distant computers and storage systems to explore, interact and understand what the data mean and to derive knowledge from them? Tools that are not only enabling but also intuitive and adaptive will be created to directly answer these needs. We have assembled a team comprised of leaders in computer science, neuroscience, genetics and knowledge discovery from the University of Southern California, the University of Chicago, the University of Michigan, and the University of Washington. We have a mature and competent staff of software developers, testers and applied scientists. And most importantly, we have experience in the challenges of big data and have developed solutions to portions of this significant and pressing big data need in biological research. Our center, entitled Big Data for Discovery Science will lead a new paradigm for interacting with large biomedical data types and scales – from 'omes' to 'organs'.

Tools. BDDS is developing branded tools to support big data discovery. These tools are designed to work independently with other BDDS components in the BDDS Platform as well as work with existing frameworks and workflows. http://bd2k.ini.usc.edu/tools/

- Unambiguously name and identify research data products with the Minimal Variable Identifier (Minid). http://bd2k.ini.usc.edu/tools/minid/
- Assemble large and complex datasets with BD Bag. http://bd2k.ini.usc.edu/tools/bdbag/
- Reduce overhead and complexity of creating and managing complex, big datasets with the Discovery Environment for Relational Information and Versioned Assets (DERVIA) http://bd2k.ini.usc.edu/tools/dervia/
- Integrate, merge and analyze large, incongruent datasets with Data Dashboard. http://bd2k.ini.usc.edu/tools/big-data-dashboard/

Findings. BDDS biomedical and computer science researchers, working in concert, have conducted studies with significant findings. A sampling includes:

- Women statin users had a 69% increased risk for conversion to mild cognitive impairment (MCI) or Alzheimer's disease (AD) when compared to women non-users. Paper submitted, New England Journal of Medicine.
- Contrary to long-standing views, men and women with the APOE ε3/ε4 genotype have nearly the same odds of developing AD from 55 to 85 years of age but women have an increased risk at younger ages (between 65 and 75). Paper submitted, Journal of the American Medical Association.
- Evaluation and confirmation of the accuracy of machine-learning methods in the classification and prediction of Parkinson's disease. Paper published in PLOS One. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157077
- The Catechol-O-methyltransferase (COMT) gene influences brain morphology and related maturation in regions mediating cognitive functions, motor control and emotional processing in neurodevelopment. Paper submitted, Nature Neuroscience.
- Presentation on the exchange of large, complex datasets to the IEEE BigData Conference, 2016. http://bd2k.ini.usc.edu/pdf/publications/bagminid.pdf

BIG DATA *for* DISCOVERY SCIENCE

Biomedical scientists now have available massive, complex datasets drawn from multiple sources generated by technologies capable of efficiently measuring biological processes with increasingly refined levels of precision. The true and as-yet unrealized potential value of these big data are limited by the scientists' ability to analyze collectively all the variables and samples to yield biological insights.

Computational methods known as causal discovery algorithms can be used to discover causal relationships from a combination of observational data, experimental data, and prior knowledge. The application of such algorithms to big biomedical data generates graphical models known as causal Bayesian networks. These networks can suggest new and important causal relationships that biomedical scientists can in turn evaluate experimentally; the networks can also reveal the presence of latent variables that scientists may choose to investigate further.

The Center for Causal Discovery (CCD) is scaling up these methods for use with big biomedical data and making them accessible to the broader biomedical research community through user-friendly tools, along with training in their application and implementation on the desktop, through the web, on a high performance computing cluster, and in the cloud.

CCD tools can be used efficiently with big biomedical datasets containing thousands or even millions of variables from thousands of samples. For example, our parallelized Fast Greedy Equivalent Search (FGES) algorithm can discover causal networks containing 30,000 variables in less than 3 minutes with high precision and recall. Unlike standard machine learning methods, which predict the value of a variable X from the observation of other variables, causal algorithms can predict the value of X from the manipulation of other variables; this capability allows a scientist to consider how a change in one or more of the other 30,000 variables might influence X and to design experiments to test hypotheses suggested by the causal network.

Although our causal discovery tools can be applied to any combination of biomedical, clinical, and other types of data, we have used three well-recognized problems both to drive algorithm and software development and to test the application of our tools in answering real-world questions:

- We applied a variation of the FGES algorithm to fMRI data to identify how specific regions of the brain influence each another. Each fMRI scan consists of a time series of measurements on ~2 mm 3 voxels, which corresponds to approximately 51,000 variables. To our knowledge, our model is currently the most comprehensive functional map of the human resting-state brain. We are now investigating feedback relations in the resting state and identifying functional connections. Our approach to discovering fine-grained differences in brain functional connectivity could be used to distinguish normal from abnormal fMRI scans and to more precisely sub-classify (and hence eventually to more effectively care for) individuals with autism spectrum disorder and other neurological conditions.
- We developed a graph algorithm for solving the causal nested effects problem and used it to discover a signaling complex of proteins coded by TP53, PTK2, YWHAZ, and MED1. This discovery guided the design of experiments which revealed that disrupting the complex blocked the transmission of aberrant signals originating from mutated TP53. We are now working with collaborators to search for drugs to target the complex. Since TP53 is mutated in about half of human cancers, the ability to block the effects of this mutation could have significant therapeutic impact.
- We developed a mixed graphical model that can discover causal relationships between both continuous and discrete variables, as well as methods for the extraction of tissue spatial heterogeneity patterns observed in histopathology images. Our approach, which can be applied to any organ and any complex disease process, allows the integration of in situ transcriptomic profiles with structural data from tissue histology slides and patient clinical data from the electronic health records to identify causal relations that influence progression in chronic diseases. We used our causal discovery methods to learn how idiopathic pulmonary fibrosis progresses spatiotemporally in the lung and which factors influence the longitudinal lung-function decline in chronic obstructive pulmonary disease. We are currently generating specific hypotheses regarding local disease progression, which we will investigate experimentally.

CCD algorithms are incorporated in a suite of software tools (http://www.ccd.pitt.edu/tools), which are open source and free. We provide a help desk, online tutorials, and in-person training on using causal discovery broadly and our tools specifically. In partnership with Harvard, we developed the means to access data, apply our tools, and share the causal networks generated securely in a cloud environment, toward facilitating their use in the NIH Data Commons.

CEDAR, with its partners at Stanford, Yale, Oxford, and Northrop Grumman, is working to transform the national ecosystem for accessing, exploring, and reusing biomedical data online. Our conviction is that the goals of open science will not be achieved unless online data sets can be annotated with metadata that are comprehensive and that reflect community-based standards. CEDAR is building new Web-based technology to make it easier for biomedical scientists to author detailed metadata that describe their experiments completely, adhere to appropriate community-based standards, and incorporate controlled terms that facilitate interoperability with other online data sets.

In the fall of 2016, our team released the first public version of the *CEDAR Workbench*. With this infrastructure, individual scientists or curators (or groups of users) can manage libraries of standard templates for defining metadata, where each metadata template is linked to a specific type of biomedical investigation and, where possible, it conforms to a particular community-based standard. The Workbench makes it easy to fill in these templates and to upload the completed metadata specifications to public or private repositories. The CEDAR Workbench makes the authoring of appropriate metadata a manageable task in a variety of ways. For example, using links to the BioPortal repository of biomedical ontologies, The CEDAR Workbench generates drop-down lists of controlled terms for filling in many template fields. And after analyzing previously entered metadata records, the software can suggest to the user possible ways to fill in certain template fields, thus streamlining the authoring process and enhancing the quality of the final specifications.

Large numbers of biomedical scientists are now eager to help us evaluate the CEDAR Workbench and to incorporate the technology into their own workflows. Here are a few of our most important collaborations:

- **AIRR/NCBI.** The Adaptive Immune Receptor Repertoire (AIRR) community began in 2014 as a grassroots effort to bring together investigators utilizing high-throughput repertoire sequencing (Rep-Seq). AIRR has produced its own metadata standard for depositing data across four different NCBI repositories: BioProject, BioSample, SRA, and GenBank. Unfortunately, none of the more than 30 databases managed by NCBI provides infrastructure for the curation of standardized metadata. AIRR is working with NCBI and with CEDAR on a pilot project to allow users to submit controlled metadata to the NCBI. Within a few months, the more than 800 members of the AIRR community will be able to use CEDAR technology to generate and upload Rep-Seq metadata. The NCBI is monitoring the AIRR activities to evaluate the potential use of CEDAR technology for the submission of metadata to all NCBI data repositories.
- **LINCS.** The Library of Network-Based Cellular Signatures (LINCS) is a consortium of six Data and Signature Generation Centers (DSGCs) and a Data Coordination and Integration Center (DCIC) supported by the NIH. LINCS scientists use a variety of experimental techniques to study how disruption of biological pathways at any one of their steps may affect cellular phenotypes. LINCS has developed extensive metadata standards, but the consortium has not yet automated the process of data and metadata submission to the DCIC, in part because of the complexities of getting metadata into the correct format. CEDAR and the LINCS DCIC are collaborating to enable all six of the LINCS DSGCs to use the *CEDAR Workbench* for metadata management, standardization, and submission to the LINCS DCIC.
- **ImmPort.** The NIAID Division of Allergy, Immunity, and Transplantation (DAIT) mandates that all its supported projects store their data in the online repository known as ImmPort. The ImmPort development team at Northop Grumman collaborates directly with CEDAR, and is pursuing using the *CEDAR Workbench* for authoring the metadata associated with ImmPort-related data sets. Our ties to the Standards Working Group of the Human Immunology Project Consortium (HIPC) give the CEDAR development team a direct link to the committee that develops metadata standards for one of the most prominent groups of ImmPort users.
- **Hydra.** Dozens of the nation's large research universities use a common platform, known as Hydra, to provide a local, public repository for open scientific data. Currently, Stanford, the Digital Public Library of America, and DuraSpace lead a major effort to extend the capabilities of the Hydra and to ease its installation and management. This work includes evaluating integration of the platform with the *CEDAR Workbench* to enable metadata authoring by all Hydra users.

CEDAR continues to receive requests from many colleagues to collaborate on the general problem of metadata authoring. Recent RFAs from the NIH relating to metadata standards and data curation have heightened CEDAR's visibility in the scientific community. Our use of metadata templates based on community-based standards and our close relationship with the BioSharing resource also enhance our visibility, and in turn link CEDAR to the ELIXIR initiative in Europe.

Characterizing Preclinical Brain Changes Leading to Alzheimer's Disease. The Center for Predictive Computational Phenotyping (CPCP) is developing innovative computational and statistical methods and software for a broad range of computational phenotyping tasks. These are tasks in which computational methods are required either to extract relevant phenotypes from complex data sources or to predict clinically important phenotypes before they are exhibited. One of the projects in the Center for Predictive Computational Phenotyping is focused on developing methods to characterize preclinical changes in longitudinally acquired brain images and other key biomarkers in order to predict cognitive decline in individuals at risk for Alzheimer's disease (AD). By the time an individual exhibits symptomatic dementia as a result of Alzheimer's disease (AD), devastating neural loss has already occurred and the window for effective intervention has already passed. CPCP has developed several novel methods that have advanced the state of the art in detecting subtle preclinical changes that are associated with AD.

While amyloid accumulation is a primary pathological event in AD, loss of connectivity between brain regions is suspected of contributing to cognitive decline. Prior studies have failed to show a close association between amyloid deposition and structural brain changes, especially in preclinical AD. Recently, we developed a novel, multi-resolution approach based on wavelets that is able to detect the influence of amyloid burden on structural brain connectivity, even in middle-aged adults, in many regions that have been implicated as important in Mild Cognitive Impairment (a precursor to AD) or full blown AD. Cognitively asymptomatic participants from the Wisconsin Registry for Alzheimer's Prevention study underwent DTI imaging to assess structural connectivity and PIB-PET imaging to measure amyloid accumulation (Fig 1a). Connectivity strengths were indexed by mean fractional anisotropy along tracts connecting 162 gray matter regions. We derived wavelet-based multi-resolution connectivity signatures for each connection and modeled its relationship with amyloid accumulation in 16 regions that accumulate amyloid plaque in AD. Using our method, we were are able to detect 25 statistically significant (10 of 25 with very strong evidence at the Bonferroni corrected level of 0.01) associations between PIB regions of interest and connectivity (Fig 1b). For example, we found that amyloid deposition in left posterior cingulate is associated with connectivity loss between temporal regions even in this preclinical stage of AD. The standard analysis showed no associations after multiple comparisons corrections.

In another related advance, we developed the first algorithm for multivariate general linear models in which the response variable is a manifold valued measurement (i.e., these are natural representations of brain changes between two visits). This approach has been critical in characterizing morphometric brain changes calculated from longitudinal T1-weighted MRI from subjects at risk for Alzheimer's disease. Our analysis shows a spatial distribution of morphometric changes strongly associated with measurable markers of AD pathology such as amyloid burden captured via tau imaging (Fig 1c). These associations were only weakly identifiable with prior methods that were based on univariate voxel-wise summaries to characterize longitudinal changes.

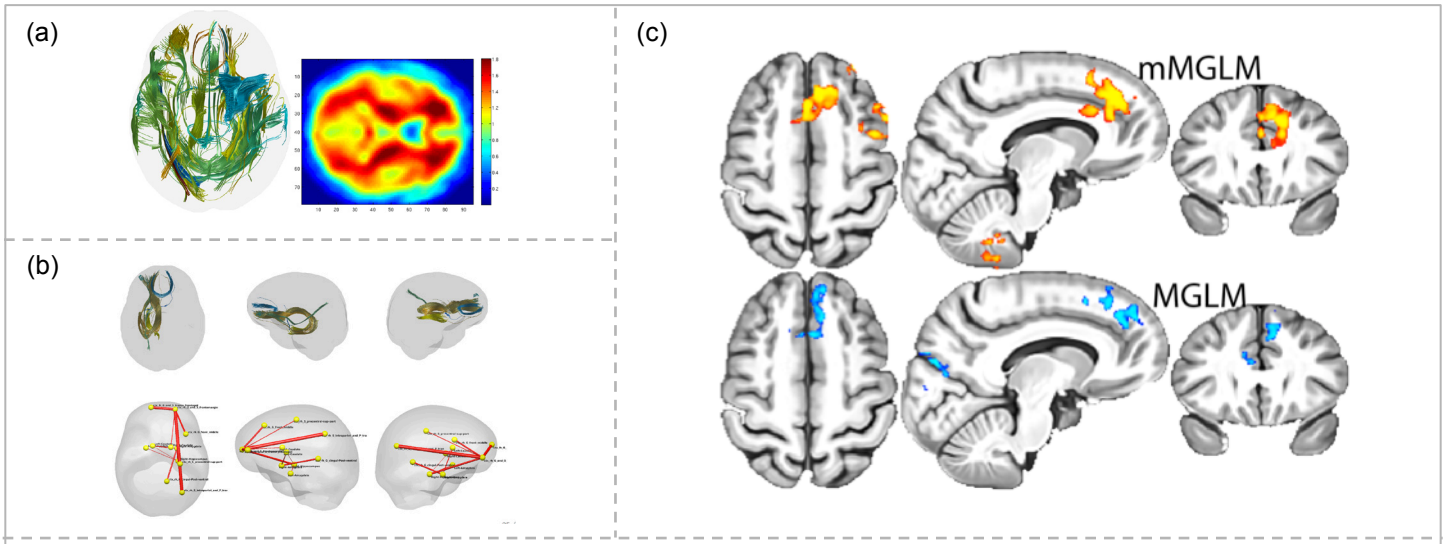
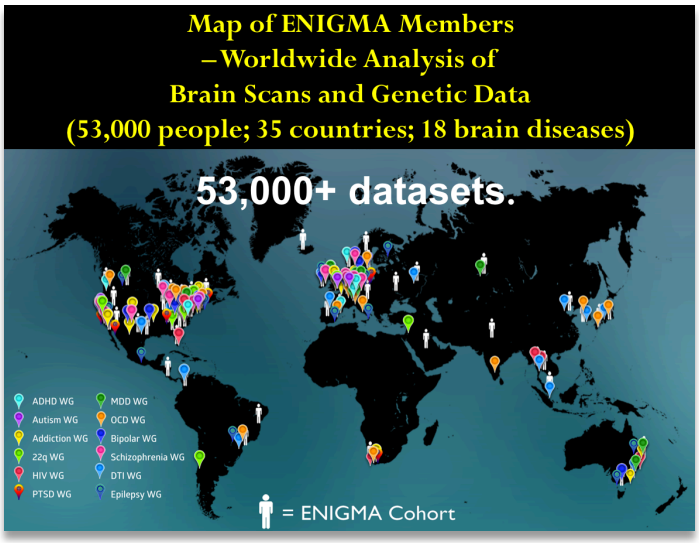
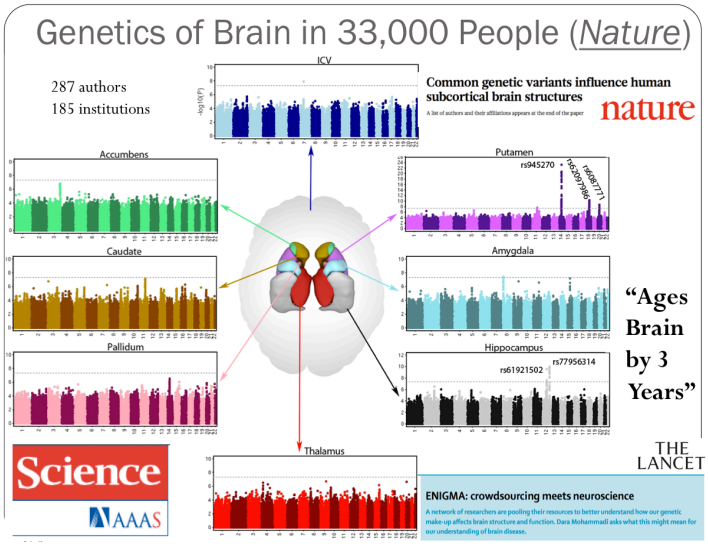


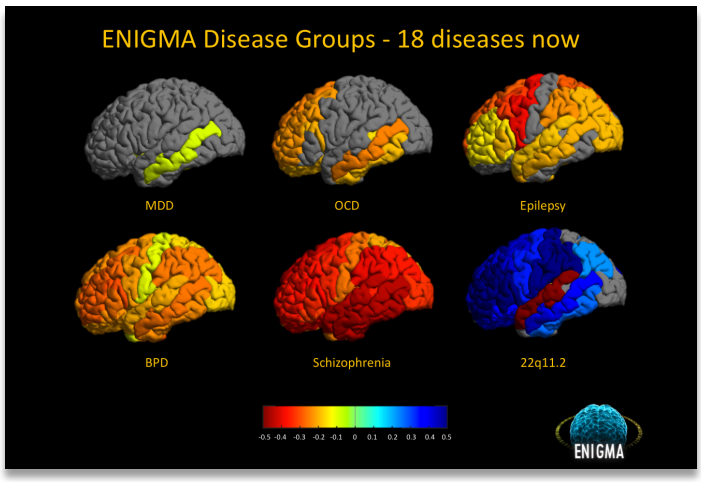
Fig 1. (a) Left: Structural connectivity from DTI; Right: Beta-amyloid burden from PIB-PET imaging. **(b)** Detected population-level connections associated with specific gray matter PIB regions of interest. **(c)** Differences in morphometric changes between individuals with amyloid+ and amyloid- pathology in healthy individuals at risk for AD. Above: differences identified by our algorithm (mMGLM) on manifold valued measurements of morphometric changes. Below: differences detected by prior state of the art on univariate summaries of changes.

Center impact. ENIGMA performs the *largest-ever studies of the human brain* – analyzing brain MRI scans from >53,000 people across 35 countries (see map below). Highlighted by the *New York Times*, *Science*, and *The Lancet* ("Crowdsourcing meets Neuroscience"), ENIGMA's 33 working groups study 18 brain diseases, uniting data, resources and talents of 700 scientists from 340 institutions.





Nature Neuroscience, *Nature*, and *Nature Communications* published ENIGMA's series of studies, "Cracking the brain's genetic code" (see above right). Building on ENIGMA's 300-author paper in Nature (Hibar 2015), 340 institutions pooled their DNA and MRI data to identify 10 genomic loci that drive brain structure and our risk for Parkinson's disease, as well as mental illness, using massively parallel distributed big data computing (Medland *Nature Neuroscience* 2015, Franke *Nature Neuroscience*). ENIGMA's two papers at *Nature Communications* map gene effects throughout the brain worldwide (Roshupkin 2016, Hibar 2017). ENIGMA's worldwide map shows how the gene *APOE* elevates Alzheimer risk throughout life and worldwide in 32,000 people scanned with MRI. ENIGMA's workshops, plenary lectures, and training events at the NIH (2), and across the EU, Russia, Siberia, the Middle East, Mongolia, the Thai Red Cross, Korea (KAIST), and Ecuador, drew thousands of attendees in aggregate, including a keynote ENIGMA lecture to thousands of radiologists at ISMRM (Toronto), the Chinese Congress of Radiology, and Russian Academy of Sciences. Lectures and tutorials are online at the ENIGMA website, with our codebases (https://github.com/ENIGMA-git/ENIGMA); ENIGMA's 10K-in-1day event in Utrecht was the largest ever supercomputing analysis of brain connectivity, analyzing connectomes from 15,000 people worldwide.



ENIGMA published the world's largest neuroimaging studies of **schizophrenia, major depression, bipolar illness, and obsessive compulsive disorder** combining MRI data from >20,000 people. Frontal abnormalities in bipolar illness contrasted with limbic abnormalities in depression, with different profiles in teenagers and adults. 3 papers in *Molecular Psychiatry* (Schmaal 2015, van Erp 2015, Hibar 2016) were covered in the worldwide media.

Introduction. HeartBD2K's primary mission is to create a user-centric, community-building, and publicly driven model platform for translating data to knowledge and advancing our understanding of health and disease. We achieve this through an interactive development process that harnesses efforts of a diverse, self-propagating user base. The premier quality of datasets, software and analytical tools attracts users, and from that community of users new data contributors and users emerge. Examples of our tools include machine learning-based text mining tools that create computable information from unstructured data (text information), identifying hidden relationships, providing community access, and contributing novel mechanistic insights to knowledgebases. Moreover, our OmicsDI and Aztec projects make use of unstructured PubMed data by retrieving metadata, thereby adding value to existing datasets or tools and ensuring their findability, accessibility, and interoperability for the entire community. Finally, our crowdsourcing platform (e.g., Mark2Cure) aids researchers in finding new information and facilitating discoveries. As the future success of any bioinformatics platform relies on the degree of adoption and support from the global population, our Heart BD2K Center places a high priority on rallying user support and enthusiasm for our digital products. We envision that a community-driven BD2K is key for reshaping the future of Big Data science.

I. Fountain of Youth Projects: *Creating sustainable value for our digital objects – extending the lifespan of data and tools*. We have made major progress in developing tools that ensure the compliance of digital objects (data & tools) with FAIR Principles. The 'Omics' Dataset Discovery Index (*OmicsDI*, https://www.omicsdi.org) provides an interface for accessing and locating datasets in multiple, globally distributed 'omics' repositories (Nature Biotechnology 2017). Meanwhile, our *Aztec* (https://aztec.bio) project enables users to access over 10,000 tools and computational resources, employing natural language processing for identifying hidden and relevant publications, and subsequently clustering, organizing, managing, as well as auto-extracting metadata from unstructured text data of PubMed, thereby rendering the tools findable and accessible for the entire community (Figure below).

II. Crowdsourcing Projects: *Leveraging community intelligence to create new biomedical knowledge*. Looking in-depth at one of our projects, the *Mark2Cure* (https://mark2cure.org) initiative both addresses and capitalizes on the exponential growth rate of scientific literature – two new articles published every minute – and allows researchers to more efficiently find information and identify new cures more rapidly. *Mark2Cure* works bilaterally in that it teaches citizen scientists to identify concepts and conceptual relationships within biomedical texts, all while utilizing advanced statistical algorithms on citizen science data in order to aid researchers in finding information within the vast ocean of biomedical knowledge. In fact, *Mark2Cure*'s first academic paper was published just less than a month ago. One specific *Mark2Cure* success involves the focused organization of biomedical literature centralized on an ultra-rare disease known as N-Glycanase 1 (NGLY1) deficiency. In collaboration with the NGLY1 Foundation and NGLY1.org, *Mark2Cure* has deployed the Center's citizen scientists to better organize knowledge of related genes, drugs, and diseases, culminating in the identification of 10,000 documents pertaining to the disease.

III. Machine Learning in Cardiovascular Medicine. The HeartBD2K team at UCLA has established an outstanding partnership with the KnowEnG BD2K Center at the University of Illinois, Urbana/Champaign, to develop cutting-edge, machine learning-based text mining and network-embedding tools. The teams have combined the phrase mining algorithms *SegPhrase+* and *ToPMine*, developed by Professor Jiawei Han at KnowEnG, with a newly developed phrase-based network-embedding technique, Large-scale Information Network Embedding (LINE). These techniques have resulted in uncovering novel patterns within 8,368 proteins relevant to cardiovascular disease; these span across six main categories of heart disease and were evaluated from 551,358 publications within the MEDLINE database between the years of 1995 and 2016. This synergistic collaboration has highly benefited both teams, in that it has accelerated the discovery of novel insights into cardiovascular disease pathogenesis, as well as achieved heretofore unattainable new developments in both machine learning and text-mining approaches from the underutilized data source of cardiovascular scientific literature.

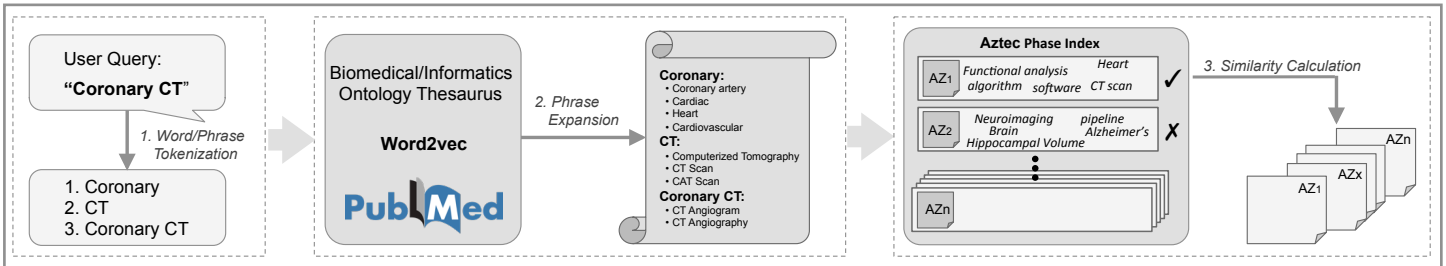


Fig. 1. Aztec Information Retrieval Workflow is illustrated here by a use case of querying "Coronary CT" (manuscript submitted).

U54 GM114838: KnowEnG, a Scalable Knowledge Engine for Large-Scale Genomic Data

PD: Jiawei Han (Univ. of Illinois at Urbana-Champaign)
Co-PD: Saurabh Sinha (Univ. of Illinois at Urbana-Champaign)
MPI: Jun S. Song (Univ. of Illinois at Urbana-Champaign) & Richard M. Weinshilboum (Mayo Clinic)

Highlight 1: Cloud-based Pipelines for Knowledge-guided Analysis of Genomics Data. We have created a Cloud-based infrastructure called KnowEnG ('Knowledge Engine for Genomics') for knowledge-guided analysis of genomics data. The user uploads their data in the form of a spreadsheet to the KnowEnG interface, and the system performs powerful data mining and machine learning tasks on those data. The unique part of such analysis is that it is carried out while making intelligent use of prior knowledge in the public domain. Such prior knowledge is represented in the form of a massive heterogeneous network called the Knowledge Network, which aggregates information from nearly 100 externally curated databases. The user may choose from several analysis pipelines (Fig. 1) to deploy on their data, including:

1. Gene prioritization. Identify genes most likely to be associated with a phenotype.
2. Phenotype prediction. Train a model to predict a numeric phenotype value from an omics profile.
3. Sample clustering. Find related groups within a collection of omics profiles. This may be used for patient stratification from their transcriptomics or somatic mutation profile.
4. Gene set characterization. Identify shared properties in a previously identified set of genes.

Each pipeline is a complex workflow involving one or more algorithms for data processing and normalization, application of the core machine learning or statistical algorithm, as well as post-processing and visualization. We are working towards integrating the KnowEnG system with other major Cloud-based data repositories such as TCGA and LINCS as part of a bigger ecosystem under the Commons umbrella (Fig. 2).

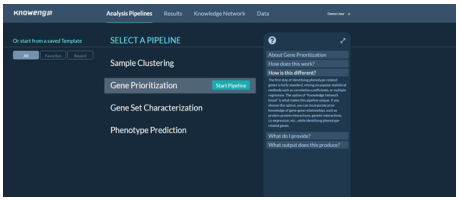


Fig 1. KnowEnG pipelines interface.

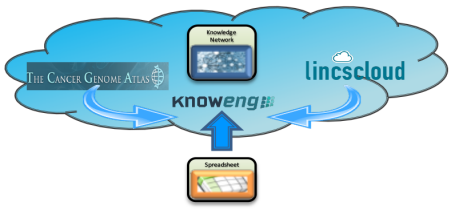


Fig 2. KnowEnG as part of the future Commons.

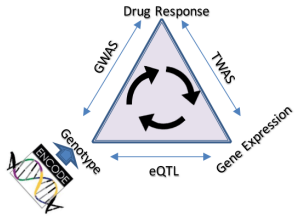


Fig 3. Integration of omics data.

Highlight 2: Advanced analytics for cancer pharmacogenomics. We have used KnowEnG analysis pipelines to rank and prioritize pathways, regulatory proteins called transcription factors (TFs), and genes associated with individual variation in drug response. Three novel algorithms have been developed to achieve this. The NetPath algorithm employs a network-based dimensionality reduction technique called Diffusion Component Analysis (DCA) to identify pathways that are most closely related to genes whose expression levels are correlated with drug response. The ProGeni algorithm employs Random Walks with Restarts (RWR) to rank genes by their association with drug response variation. The pGENMi algorithm uses Probabilistic Graphical Models (PGM) to integrate genotype (SNP), DNA methylation, gene expression, and phenotype measurements on a panel of cell lines, along with prior knowledge of the regulatory genome (ENCODE project), to identify transcription factors likely to influence drug sensitivity (Fig. 3). We have experimentally validated 23 predictions of genes associated with specific drugs.

Highlight 3: Literature mining for disease-specific proteins. Typical text mining tools take substantial human efforts for manual data curation and extraction of structure from text data. In contrast, we have been developing a data-driven, semi-supervised text mining paradigm to mine a massive collection of biomedical texts. To ensure such effort will benefit biomedical research, we have been collaborating with the HeartBD2K Center at UCLA to conduct biomedical literature mining for disease-specific proteins. We analyzed hundreds of thousands of research abstracts in PubMed related to cardiovascular diseases in the last 20 years, comparatively ranking 250 given proteins with respect to each of six subcategories of cardiovascular disease. The results reveal new insights on disease-specific proteins. We are further proceeding to uncover distinctive dense protein subnetworks for the subcategories of heart diseases. We are also developing scalable literature retrieval functions for effectively finding (ranking) research papers given a *set* of biological entities (such as a *set* of genes). These biological text mining functions will be built into the KnowEnG engine and made accessible to the general biomedical research community.

SegPhrase+ https://github.com/shangjingbo1226/SegPhrase
ToPMine https://github.com/ydj0604/ToPMine-Spark

In a unique project, researchers from the BD2K-LINCS Data Coordination and Integration Center have crowdsourced the annotation and analysis of a large number of gene expression profiles from the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO). More than 70 volunteers from 25 countries helped analyze the data, enabling the identification of new associations between genes, diseases, and drugs – something that a smaller number of unaided researchers, or an automated computer program, would not be able to achieve. An article published in the journal *Nature Communications* describes the crowdsourcing project. Omics repositories, which are virtual storehouses for raw gene expression data, contain thousands of studies. Such an abundance of data opens opportunities for integrative analyses that can uncover new knowledge that was missed or was not possible in the initial publication of the data. For example, while a dataset from a given study may have been used for a particular published article, that same dataset may contain evidence whose value can only become realized when combined with data from another study. Then, it might become apparent that a drug can be repurposed to treat a different disease. Several computerized search engines have been designed to comb through this data. However, for these tools to be effective, they require heavy, time-consuming human curation to ensure accuracy. That is where crowdsourcing can be useful. For this project, the 70 volunteers were recruited through a massive open online course (MOOC), which was being taught on the Coursera MOOC platform by the BD2K-LINCS DCIC. The student volunteers were asked first to identify relevant studies in the NCBI GEO database – in this case, studies in which single-gene or single-drug perturbations were applied to mammalian cells, or in which normal versus diseased tissues were compared. Once the studies were selected, the volunteers extracted metadata from the studies and then computed differential expression using a custom-designed Chrome browser extension developed by the BD2K-LINCS DCIC.

This process extracted information about gene signatures – observations of groups of genes whose combined expression is associated with a particular condition or drug action – which were stored in a new database. Then, the BD2K-LINCS DCIC used the database to visualize and analyze the signatures on a web portal known as Crowd Extracted Expression of Differential Signatures, or CREEDS, which was developed by the BD2K-LINCS DCIC. Over the course of the project, over 3,100 single-gene perturbations from more than 2,300 studies were submitted, as well as 1,238 single-drug perturbations from nearly 450 studies. By utilizing volunteers, so called 'citizen-scientists,' the BD2K-LINCS DCIC was able to bring a much greater scale of human curation and quality control than could be by a single group. Ultimately, the manually extracted signatures were used as a training set to help a program that uses machine learning to process all the data currently available in GEO for adding more drug, gene, and disease signatures to the CREEDS database. While researchers generally find that the quality of automatically generated signatures is subpar compared to those created by humans, such crowdsourced efforts can be integrated with machine learning to help refine the data. Instances that the computer programs find more difficult can be presented to the crowdsourced human curators with suggestions; this allows for higher quality data, while reducing effort required of the volunteer. While many new relationships between genes, drugs, and diseases were identified, further hypotheses can be formed through additional analysis of the data. The BD2K-LINCS DCIC has made the data and analysis of it available to the public on the CREEDS portal. To interact with the portal, visit http://amp.pharm.mssm.edu/creeds.

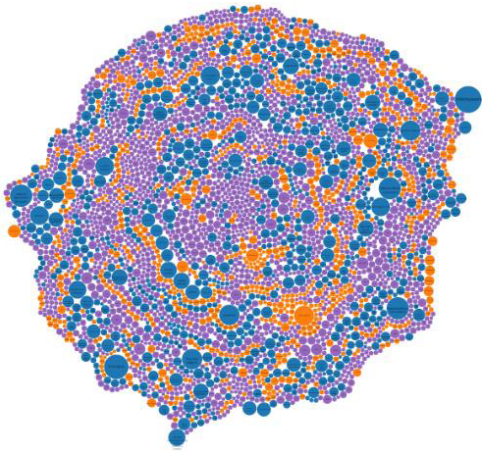


Fig 1. Similarity between drug, gene and disease perturbation gene-expression signatures visualized as an interactive web-based bubble plot where each node represents a drug, a gene or a disease.

Enabling High-frequency Mobile Sensor Data Collection for Development and Validation of Novel Multi-sensory Biomarkers and Sensor-triggered Interventions

Biomedical research studies archive biospecimens in biobanks so that the biospecimens can be reprocessed to take advantage of future improvements in assays and support biomedical discoveries not possible at the time of data collection. Mobile health (mHealth) studies, on the other hand, usually collect digital biomarkers (e.g., activity counts) that are specific to the computational models adopted by respective vendors at the time of data collection. This approach prevents any future validation of these biomarkers and makes it impossible to recompute newer biomarkers. To obtain a similar, long-lasting research utility as the biobanks, raw sensor data must be collected in a way that allows it to be reprocessed in the future to validate prior biomarkers and to obtain new biomarkers. In addition, data science and computational research for the development and validation of new biomarkers needs to collect raw sensor data and associated labels.

The MD2K Center of Excellence has developed and released an open-source software suite (http://github.com/MD2Korg/) called *mCerebrum* that is designed from the ground up as a high-frequency data stream processing tool chain. It supports concurrent collection of streaming data from multiple sensors in phones and wearables, including wrist- and chest-worn sensors, smart toothbrushes, as well as weight and blood pressure monitors. mCerebrum supports high-frequency raw sensor data collection in excess of 70+ million samples/day, along with their curation, analysis, storage (2GB/day), and secure upload to a cloud storage. Built-in privacy controls allow participants to suspend/resume data collection from specific sensors and to control notifications/interruptions.

Data science research conducted by MD2K has already resulted in ten mHealth biomarkers, including stress, smoking, craving, eating, activity, and drug use. The entire pipeline of mobile sensor big data – collection, curation, feature extraction, biomarker computation, time series pattern mining, and micro-randomization – has been developed and fully-implemented on the phone to support real-time, biomarker-triggered notifications and interventions.

mCerebrum provides native support for triggering notifications, self-report prompts, and interventions based on real-time values of digital biomarkers derived from sensor data. To collect momentary self-reports, participants can be prompted randomly based on time of day, based on self-reported events, and now also based on events detected by sensors. All prompts, notifications, and interventions are carefully coordinated to limit the burden on participants while meeting study requirements. There are a variety of ways a user can customize mCerebrum, including the applications, sensors, storage locations, questionnaires, notifications, and scheduling algorithms. mCerebrum provides more than 25 unique implementations and allows a user or study coordinator to select from and configure them. This ease of configurability has allowed mCerebrum to be deployed in seven field studies (smoking, eating, oral health, cocaine use, and congestive heart failure) being conducted at seven unique sites throughout the United States. These studies will result in a total of 584,640 hours of high-frequency sensor data, consisting of more than 2.2 trillion data points, for a total of at least 146 TB. These studies support projects from NIBIB, NIDA, NCI, NIMH, and NIDCR from NIH, as well as NSF.

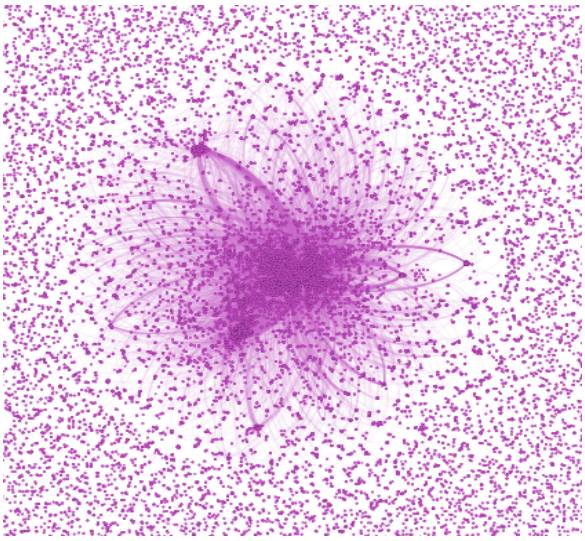
About MD2K. The MD2K Center brings together investigators, students, and postdocs in Computer Science, Engineering, Medicine, Behavioral Science, and Statistics, drawn from 12 universities (Cornell Tech, Georgia Tech, Northwestern, Ohio State, UCLA, UC San Diego, UC San Francisco, the University of Massachusetts Amherst, the University of Memphis, the University of Michigan, the University of Utah, and West Virginia University) and Open mHealth (a non-profit organization). The MD2K Team is developing innovative tools to make it easier to gather, analyze, and interpret health data generated by mobile and wearable sensors. The goal of the big data solutions being developed by MD2K is to reliably quantify physical, biological, behavioral, social, and environmental factors that contribute to health and disease risk. The research conducted by MD2K is expected to improve the health of individuals through early detection of adverse health events and by facilitating prevention. The MD2K team is directly targeting two complex health conditions with high mortality risk – reducing hospital readmission in congestive heart failure (CHF) patients and preventing relapse in abstinent smokers. The approach and product of MD2K is also applicable to other complex diseases, such as asthma, substance abuse, and obesity. The Center has made the MD2K tools, software, tutorials, videos, and other training materials widely available. It regularly organizes webinars to encourage their use by researchers and clinicians.

Planetary-Scale Mobile Health Data Reveal Factors that Affect Physical Activity

Summary. The rise of consumer fitness trackers, smartphones, and smart watches offers the opportunity for an unprecedented examination of physical activity levels and motivation. The Mobilize Center, a BD2K Center of Excellence at Stanford University, is analyzing movement data from 6 million individuals using a smartphone app for activity and health tracking. Their studies reveal new insights about physical activity levels around the world and what factors are predictive of these activity levels. Further, the work provides a new methodological framework for studying health behaviors, like physical activity, on a global scale, furthering the BD2K goal to "enhance the research community's capabilities for using Big Data in biomedical research."[1]

Background. Physical inactivity contributes to ~5 million deaths per year worldwide,[2] yet little is known about activity dynamics throughout a day for countries around the world and more data is needed to understand what environmental, social, and policy factors increase physical activity. Typical studies utilize relatively small sample sizes and surveys, whose results are often distorted (e.g., due to poor memory), limiting our understanding. The Mobilize Center has analyzed a dataset of 68 million days of activity for 6 million individuals from 100+ countries who use the Argus smartphone app from Azumio for activity and health tracking. The dataset is three orders of magnitude larger than previous studies of physical activity and objectively measures physical activity in a free-living environment.

Findings. We have used the data to develop new metrics to characterize activity in a population. These metrics are helping us to understand the relationships between activity and obesity levels in different subgroups of the population (e.g., age and body-mass index groups) and to identify the factors that predict who will be most and least active. For example, our analysis shows a statistically significant increase in physical activity for those who participate in the app's social network, an effect that lasts for 3 months after joining the network and is greater for obese individuals (body-mass index [BMI] >= 30) than those of normal weight (18.5 ≤ BMI< 25), but only if the individual receives a friend request. Among those who send a friend request, the change in physical activity levels is greatest for those of normal weight.[3] Those with many social connections (*center of image shown on the right*) are more active than those with few or no connections. Interestingly, the first connection to the social network has the greatest impact on user behavior, compared with subsequent connections.



Impact and future work. The Mobilize Center has delivered on the promise of utilizing data from physical activity-tracking devices to reveal worldwide patterns of activity and provide new insights into the key factors that influence activity levels. Armed with this knowledge, policy makers, clinicians, and others can more effectively design targeted interventions that increase activity and curb the obesity epidemic. The Mobilize Center has already demonstrated the potential for doing this, developing a predictive model that identifies those most likely to benefit from joining Argus's social network.[3] This work is but one example of how the Mobilize Center has brought world-class computer and data scientists into the biomedical workforce. The software for analyzing this data will be open-source, like all the other tools disseminated by the Mobilize Center[4] which are already accelerating the research of thousands of individuals world-wide.

[1] https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-009.html.
[2] Lee I, *et al*., *The Lancet*, 2012, 380(9838), 219-229.
[3] Althoff T, *et al.,* to be presented at WSDM, Feb. 2017, https://arxiv.org/abs/1612.03053.
[4] Here are the major tools:
Snorkel: https://github.com/HazyResearch/snorkel
DeepDive: http://deepdive.stanford.edu/
SNAP: https://snap.stanford.edu/index.html
OpenSim: http://opensim.stanford.edu/
See http://mobilize.stanford.edu/software-and-data/ for a complete listing.

Creating a standard programmatic interface for genomic data with the Global Alliance API. The Global Alliance Genomics API allows for the interoperable exchange of genomic information across multiple organizations and on multiple platforms. This is a freely available open standard for interoperability that uses common web protocols to support serving and sharing of data on DNA sequences and genomic variation. It overcomes the barriers of incompatible infrastructure between organizations and institutions to enable DNA data providers and consumers to better share genomic data and work together on a global scale, advancing genome research and clinical application. The APIs were contributed by 15 different task teams of the Global Alliance and 100 community contributors. The codebase has been forked 230 times. The adoption of the API is growing, with major efforts coming from UCSC, Ensembl, and Google. Other institutes are using the data definitions to create Global Alliance interfaces, including Washington University, Microsoft, Cornell, and UC Berkeley. The integration team at UCSC is building an ecosystem of tools around the APIs of the open source Global Alliance reference implementation server, including a suite of test tools to verify compliance with the Global Alliance data definitions, a mature python client library, interactive API documentation using Swagger, a Dockerization of the reference server, and many more utilities to facilitate adoption of this standard.

BRCA exchange. Roughly 60% of women who inherit a pathogenic BRCA variant will develop breast cancer by the age of 70, versus 12% of all women. Further, as many as 39% of women with a pathogenic BRCA variant will develop ovarian cancer, vs. 1.3% of all women. This has led many women to undergo BRCA testing to understand and manage their risk. Unfortunately, many women will be informed that they carry a Variant of Unknown Significance (VUS). Research on BRCA variation was long hindered by a private gene patent. Gene patenting was struck down by the U.S. Supreme Court in 2012, but in its wake, the data on BRCA variation has remained fragmented with no single source for all available data. Consequently, most doctors and geneticists are working with incomplete information. To address this, the Global Alliance launched the BRCA Challenge, a consortium with a goal of cataloguing all public knowledge on BRCA variation. Through this effort, we developed BRCA Exchange (http://www.brcaexchange.org), which aggregates and unifies knowledge from genetic repositories including ClinVar, LOVD, 1000 Genomes, ExAC, BIC, exLOVD, ESP, and ENIGMA. BRCA Exchange is currently the largest public repository of BRCA variation, with over 17,700 variants. These data are publicly available to download and query via the Global Alliance API. The ENIGMA consortium is utilizing this data for expert curation of BRCA variants. Since 2016, they have tripled the number of variants that they have curated, with expert reviews of close to 3300 variants. Moving forward, we are working with ENIGMA to identify and aggregate the information needed to curate additional variants, including family history and case-level data, so that fewer women in the future will receive VUS test results, thus empowering women to better understand and manage their own hereditary risks.

Large-scale, cloud-based analysis of cancer data. Creating mobile workflows and tools designed to work across clouds is a challenge for genomics researchers. Our Center has worked closely with the Global Alliance Containers and Workflows task team in creating a standard for this approach. Central to this success is the creation of the Dockstore, our best-practice platform for exchanging workflows and tools using Docker. This approach was instrumental for projects like the ICGC's PanCancer Analysis of Whole Genomes (PCAWG), which relied on portable workflows to compute across multiple environments, ultimately peaking at 14 cloud and HPC environments around the world. Central to our large-scale use of Dockstore workflows and tools is our Toil project, which is a platform for running tools and workflows at scale on AWS and other clouds. The latest Toil recompute of 20,000+ RNASeq samples shows the power of combining Docker-based tools with a highly scalable workflow engine. The future of large-scale genomics work, using Dockstore, Toil, and related Global Alliance standards together, complements the vision of the Data Commons.

Genome graphs. Our goal has been to capture the natural diversity in the human genome sequence that exists in different ethnic populations around the world so as to make genome analysis more accurate, more complete, and no longer biased toward the particular ethnic groups who contributed to the single reference human genome we use today. This diversity is captured in a rich and beautiful mathematical structure, the Human Genome Variation Map (HGVM), called a genome graph. In collaboration with the Global Alliance, we have made significant progress constructing a standard data model, API, and software implementation for HGVM.

We have focused on the development of the basic data science for the project; however, we have constructed whole human genome graphs incorporating all variations from the 1000 Genomes Project's Phase 3 data release – some 50 million point variations. We used this with our prototype variant calling pipeline to determine variants within test samples as part of the Precision FDA evaluation. This was the first time anyone had performed end-to-end variant calling across a complete human genome using a genome graph. Although the results – for a number of anticipated technical reasons – were not yet competitive with the heavily optimized best of breed pipelines, we still view this as an important milestone in the scaling of the project. In the future, as we publish the variant calling pipeline, we anticipate refining and releasing this draft HGVM as a product of the GA4GH reference-variation consortium.

The NHANES PIC-SURE Project. The NHANES project of the CDC has been one of the most impactful public health research projects of this country. It has direct relevance to nutrition, pediatric growth, environmental exposures, and even genetic components of health risk. Since 2000, it has been updated continually and is used by multiple branches of medicine, epidemiology, public health, and policy studies. NHANES comprises measurements of over 40,000 individuals who are selected to be demographically representative by CDC-sponsored researchers. Analyzing these data presents multiple challenges. As they are currently configured, the files describing different aspects of these populations are spread over several directories, each with their own data dictionaries and data types. Bringing together the data from different studies on an individual takes considerable effort and significant time to understand how all of the data relates to individual study participants. Among the consequences of this challenge is that researchers only become familiar with a few measurement modalities in each substudy and rarely combine the rich variety of data types across all the data types. Doing so comprehensively in the omics style is even rarer. Because one of the goals of PIC-SURE is to enable the researcher/programmer/analyst to focus on their research questions rather than the computational, data merging, and data access tasks, the NHANES data sets were chosen as one the first PIC-SURE demonstration projects—referred to below as the NHANES PIC-SURE project.

Our project illustrates three of the impactful contributions that PIC-SURE brings to the community. Foremost is enabling the combination of thousands of different variables—clinical, environmental, self-reported, biochemical—all in one set of data structures that are queryable and made available as an internationally shared resource on the cloud. The second contribution is the development of a standardized web API, which allows programmers unfamiliar with the underlying data structures across all the NHANES data to issue commands using a web-borne interface that is familiar to all modern programmers (i.e., a web-API, https://github.com/hms-dbmi/IRCT-API). Third, our project illustrates how we can use these large and complex data sets in a reproducible way, enabling other researchers to follow exactly what the published analyses have been and to modify them as they wish. In so doing, our project addresses three large challenges of the biomedical big data era. Even as a demonstration, the NHANES PIC-SURE project has been successful at the international level. It was recently featured in *Nature Scientific Data* (*Sci Data* 2016;3:160096) for expanding the user community of these important data.

Similar motivations across all kinds of biomedical data have led to multiple APIs to access data. For example, there are genomics-specific APIs, EHR-specific APIs, wearable-specific (e.g., Fitbit) APIs, and APIs for environmental, geographic exposures. However, there is no single API that allows the application programmer to mix and match across different data modalities without knowing the myriad details of the data structures underlying each of the data sources. In addition to the complexity of the data structures, there is also the additional challenge of multiple alternative terminologies used even within a single data modality. Our API provides that single programmatic interface, and this is illustrated in its implementation for the NHANES API. The API and open source supporting software are publicly accessible at http://bd2k-picsure.hms.harvard.edu/. The underlying data are at https://nhanes.hms.harvard.edu/transmart/datasetExplorer/index (username/password = demo/demo, Fig. 1).

By providing access to all these different data sources in this single web API, we allow the programmer to perform arbitrary combinations of the characteristics specified across these different data modalities to retrieve patient populations meeting these characteristics across all those data modalities and their component data types. Moreover, users of the above referenced NHANES/PIC-SURE website can see our adopted standardized authentication system, which distributes the authentication and authorization process to a variety of different servers that meet the needs of each party. To address growing concerns in biomedical research about reproducibility, particularly in "big data" projects, we have engineered the standardization of the PIC-SURE API such that it can be called from a multiplicity of different programming languages and within electronic notebooks. We have specifically illustrated this using the open source Jupyter notebook, which has been widely adopted across multiple data science disciplines both for its usefulness in keeping a record of data analyses and for allowing reproducibility of the analyses, given access to these notebooks. This means that the full suite of Jupyter is available to the PIC-SURE community. Please find details at http://bd2k-picsure.hms.harvard.edu/example01.html.

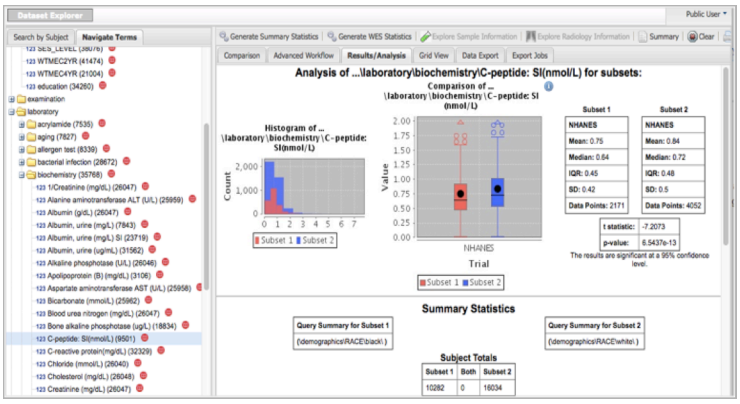


Fig 1. PIC-SURE dataset explorer presents the difference of C-peptide serum levels between Caucasians & African Americans.

In summary, we have achieved one of the major aims of the PIC-SURE Center of Excellence by providing scalable integration in a modern queryable format for web application programmers that abstracts away the complexity and lack of standardization within each of these data modalities and certainly, across them. We are currently engaged in several projects using this infrastructure beyond the NHANES study, and we are glad to partner with other investigators to enable their use of these tools.

U54HL127366: Broad Institute LINCS Center for Transcriptomics and Toxicology
PD: Todd Golub (Broad Institute),
PI: Aravind Subramanian (Broad Institute)

Introduction. The overarching goal of the Broad Institute's transcriptomics center is the development of comprehensive signatures of cellular states that can be used by the entire research community to understand protein function, small-molecule action, physiological states, and disease states. An important component of this effort is to develop the computational tools necessary to extract biological insights from the growing set of government-funded public gene expression resources by creating workflows that facilitate integrative analysis.

APIs to 2M gene expression profiles. Our center is creating the world's most comprehensive resource of perturbational signatures as part of the LINCS consortium. This already includes 2,234,115 transcriptional signatures that we have generated from genetic loss of function (CRISPR knock-out, shRNA knockdown), gain of function, and small-molecule (drug and tool compound) perturbations spanning 50 cell types of varied lineage. To facilitate the incorporation of these data with other resources created in BD2K, we have developed 19 distinct APIs that address a range of algorithmic and metadata lookup functions. For example, a user who seeks to find a drug that modulates a gene of interest can simply call an API rather than download and process the dataset. Because we generated the original dataset, we have been able to annotate these data such that other users can benefit from our experience with quality control and analysis. Already, over 50 early access developers are using these APIs for a variety of research applications.

Bringing large-scale datasets to bear on toxicology research. We are collaborating with researchers at the Environmental Protection Agency (EPA) and the National Institute of Environmental Health Sciences (NIEHS) to apply transcriptional signatures and associated metadata to inform the safety of compounds in the environment. These efforts have produced rich annotations for over 400 small molecules actively being studied by the toxicology community. One such molecule identified to have an unusually strong transcriptional effect was triclosan, which is used in consumer products but has recently come under scrutiny for its public health risk. Our analysis suggested that triclosan inhibits glycogen synthase kinase 3, and these results are being reviewed with NIEHS collaborators. Motivated by this result, we are developing a workflow to connect indications mined from Tox21 resources such that they are accessible in the same web app environment for analysis used by Connectivity Map and LINCS scientists.

Web apps at clue.io and the Drug Repurposing Hub. The dramatic increase in high-dimensional perturbational datasets available to the biomedical community has revealed the need for intuitive and performant user-interfaces to explore and query these data. We have developed a computational environment, called CLUE, to execute on state-of-the-art cloud-based systems. This environment makes data and tools available on the cloud, harmonizes datasets to facilitate interoperability between perturbational data types, and implements web applications with user friendly graphical user interfaces that access underlying sophisticated algorithms at https://clue.io. One application of this software platform is to enable drug repurposing - there is a pressing need for a comprehensive library of clinical drugs. We have created a valuable screening resource, the Broad BD2K Repurposing Hub, that enables the systematic evaluation of drug function across a variety of information-rich cellular high-throughput assays to generate repurposing hypotheses. Our effort resulted in two complementary components: 1) A best-in-class physical collection of 5,000 well-annotated compounds with more than 3,000 clinical drugs; and 2) A web application to enable browsing of library contents (including extensive drug annotations). The Hub enables active participation of the user community via drug library revisions and additions, review and annotation of connectivity reports, and deposition of new assay results at https://clue.io/repurposing.

Fig. 1. Illustrated here is the Drug Repurposing Hub, a software platform to enable the systematic cataloging and characterization of drug function by the user community. Drug will be profiled across cellular gene expression, cytotoxicity, morphology, and proteomic assays.

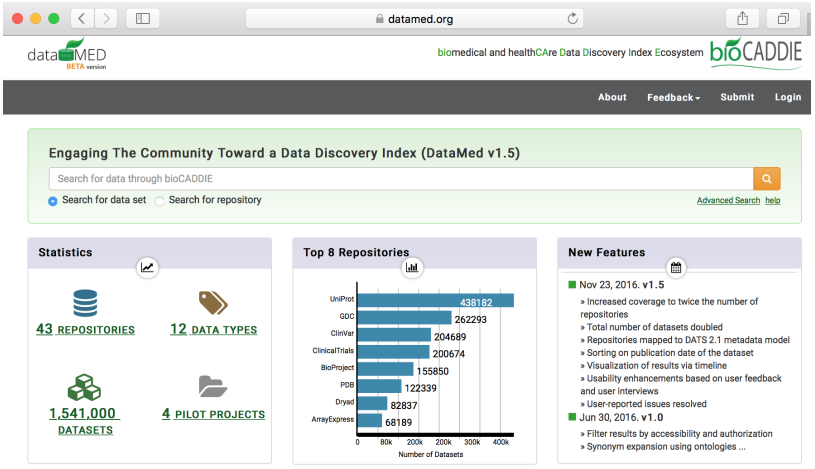
biomedical and HealthCAre Data Discovery Index Ecosystem (bioCADDIE)

PI: Lucila Ohno-Machado (UCSD), Executive Committee: George Alter (University of Michigan),
Ian Fore (Science Officer, NCI), Jeffrey Grethe (UCSD), Susanna Sansone (Oxford University), Hua Xu (UTHealth)

bioCADDIE[1] is a consortium led by the University of California San Diego and composed of University of Texas Houston, University of Michigan Ann Arbor, and Oxford University. It is charged with exploring, in a working prototype, the foundations for a Data Discovery Index (DDI); a "catalog" of data across multiple biomedical data repositories. The DDI makes data objects findable by using various search criteria, including accessibility status. Its goal is to help NIH provide a data index as part of its infrastructure and *to do for data what PubMed did for the biomedical literature*. The NIH Commons needs an organized index of digital objects and bioCADDIE is a U24 resource grant to help build such a data index. To accomplish this, bioCADDIE has to create a whole ecosystem in which all details, from object unique identifiers, to metadata specifications, to rewards for sharing in easy-to-index formats and for citing data need to be addressed. bioCADDIE is a concrete instantiation of a critical piece of the BD2K Commons.

Accomplishments to date: bioCADDIE's activities started effectively 22 months ago. Timeline and deliverables were approved by NIH on March 2015, and are available in a white paper. The project has met all its milestones so far. bioCADDIE engages the scientific community for guidance and collaboration. There are several prerequisites for building a DDI. Thirteen working groups, with more than 99 participants, and supplemental pilot projects studied critical issues and provided recommendations about object identifiers, metadata specifications, [2] use cases and benchmark materials, criteria for repository inclusion, and data citation strategies. Another way we engage the community is through solicitation, competitive review and awarding of pilot projects: bioCADDIE has funded 23 investigators outside the originally funded team, and involved 15 additional institutions (UCLA, Mayo Clinic, U Utah, L Berkeley Lab, OHSU, Stanford, Harvard, U Delaware, FORCE11, U Texas Houston, UCSD, Emory, Rutgers, NICTA, POSTECH) via pilot projects. Additionally, bioCADDIE's supplements and challenges funded 103 people in 77 institutions. These highly innovative pilots create new modules for possible inclusion in the DDI prototype (e.g., innovative ranking approaches, mining the literature for data).

DataMed is the DDI prototype developed by bioCADDIE's core development team. As of 2/1/17, it has 6,841 users. The software is open source and available in GitHub, and community feedback is used to improve it. The March version of DataMed will cover more than 60 highly utilized repositories (e.g., PDB, GEO, SRA, dbGaP, UniProt, GDC, Clinical Trials, BioProject and Dryad) and more than 1.5 million data sets.



BD2K interactions and training: bioCADDIE has so far collaborated with the following eight BD2K centers: Center for Big Data in Translational Genomics, PIC-SURE, MD2K, BD2K-LINCS DCIC, Heart BD2K, BDDS, CEDAR, CCD, in various projects. Although bioCADDIE is not required to have a training mission, its investigators have supervised 19 students and 3 postdoctoral fellows for this project. Additionally, bioCADDIE has organized 27 webinars, which have been freely viewed over 1670 times on YouTube. Its workshops were attended by over 73 people from 39 institutions.

Plans for the near future: bioCADDIE will continue to gather feedback through an established process by which community input is used to establish best practices and instantiate them in DataMed. It is a rapidly evolving project that remains responsive the other developments towards the BD2K Commons that are just starting to materialize. bioCADDIE will work with BD2K programs, NIH and the scientific community to ensure that proper APIs make the core development accessible by various applications and to enhance indexing automation. In addition to looking for input by the scientific community at large, it plans to work specifically with NLM and other institutes to be aligned with and contribute to the various indexing initiatives at NIH.

---

[1] bioCADDIE introduction: bioRxiv 094888; doi: https://doi.org/10.1101/094888 (a revised version is under review)
[2] DATS metadata specification paper: bioRxiv 103143; doi: https://doi.org/10.1101/103143 (under review)

APPENDIX – I

BD2K Centers of Excellence: A List of Software Tools Contributed Since October 2014

Name	Center	Link
Beagle	BDDS	http://sts.thss.tsinghua.edu.cn/beagle
TransProteomics Pipeline (TPP)	BDDS	https://moritz.systemsbiology.org/resources/software/
Big Data Dashboard	BDDS	https://github.com/SOCR/SOCR-Dashboard
DERVIA	BDDS	http://bd2k.ini.usc.edu/tools/dervia/
Big Data Repository (BDR)	BDDS	http://bd2k.ini.usc.edu/tools/big-data-repository/
GEM	BDDS	http://bd2k.ini.usc.edu/tools/gem/
BDDS Galaxy	BDDS	http://bd2k.ini.usc.edu/tools/association-studies/
Neuroimaging Phewas	BDDS	http://bd2k.ini.usc.edu/tools/neuroimaging-phewas/
Panther Overrepresentation Test	BDDS	http://bd2k.ini.usc.edu/tools/panther/
MINID	BDDS	http://bd2k.ini.usc.edu/tools/minid/
BDBAG	BDDS	http://bd2k.ini.usc.edu/tools/bdbag/
Big Data Quality Control (BDQC)	BDDS	http://bd2k.ini.usc.edu/tools/big-data-quality-control/
BRCA Exchange	BDTG	https://github.com/BD2KGenomics/brca-exchange
DCC Dashboard	BDTG	https://github.com/BD2KGenomics/dcc-dashboard
CGCloud	BDTG	https://github.com/BD2KGenomics/cgcloud
Toil	BDTG	http://toil.readthedocs.io/en/releases-3.5.x/
CGL Docker Lib	BDTG	https://github.com/BD2KGenomics/cgl-docker-lib
ProTECT	BDTG	https://github.com/BD2KGenomics/protect
BWA	BDTG	https://github.com/BD2KGenomics/bwa
ADAM	BDTG	https://github.com/BD2KGenomics/adam
React-Autosuggest	BDTG	https://github.com/BD2KGenomics/react-autosuggest
Progressive Cactus	BDTG	https://github.com/BD2KGenomics/cactus
HGVM-Builder	BDTG	https://github.com/BD2KGenomics/hgvm-builder
23andMe-brca-exchange	BDTG	https://github.com/BD2KGenomics/23andMe-brca-exchange
Conductor	BDTG	https://github.com/BD2KGenomics/conductor
React Data Components	BDTG	https://github.com/BD2KGenomics/react-data-components
10k Exomes	BDTG	https://github.com/BD2KGenomics/10k-exomes
sonLib	BDTG	https://github.com/BD2KGenomics/sonLib
GATK Whole Genome Pipeline	BDTG	https://github.com/BD2KGenomics/gatk-whole-genome-pipeline
GA4GH Gemini	BDTG	https://github.com/BD2KGenomics/ga4gh-gemini
Toil-VG	BDTG	https://github.com/BD2KGenomics/toil-vg
Toil-RNAseq	BDTG	https://github.com/BD2KGenomics/toil-rnaseq
Toil-TOPMed	BDTG	https://github.com/BD2KGenomics/toil-topmed
Redwood Client	BDTG	https://github.com/BD2KGenomics/dcc-redwood-client
DCC Ops	BDTG	https://github.com/BD2KGenomics/dcc-ops
S3AM	BDTG	https://github.com/BD2KGenomics/s3am
DCC Dockstore Tool Runner	BDTG	https://github.com/BD2KGenomics/dcc-dockstore-tool-runner
Spinnaker	BDTG	https://github.com/BD2KGenomics/dcc-spinnaker

Name	Center	Link
Biomedicine API Examples	BDTG	https://github.com/BD2KGenomics/bioapi-examples
DCC Storage	BDTG	https://github.com/BD2KGenomics/dcc-storage
Toil Lib	BDTG	https://github.com/BD2KGenomics/toil-lib
DCC Action Service	BDTG	https://github.com/BD2KGenomics/dcc-action-service
DCC Portal	BDTG	https://github.com/BD2KGenomics/dcc-portal
Causal Web	CCD	https://ccd2.vm.bridges.psc.edu/ccd/
Causal Command	CCD	https://internal.rods.pitt.edu/datadepot
Py Causal	CCD	https://github.com/bd2kccd/py-causal
R Causal	CCD	https://github.com/bd2kccd/r-causal
TETRAD	CCD	https://github.com/cmu-phil/tetrad
CEDAR Workbench	CEDAR	http://cedar.metadatacenter.net/
EBSeqHMM	CPCP	https://www.bioconductor.org/packages/release/bioc/html/EBSeqHMM.html
Oscope	CPCP	http://www.bioconductor.org/packages/devel/bioc/html/Oscope.html
OEFinder	CPCP	https://github.com/lengning/OEFinder
Rolemodel	CPCP	https://github.com/wiscstatman/rolemodel
GADGET	CPCP	http://gadget.biostat.wisc.edu/
rvalues	CPCP	https://cran.r-project.org/web/packages/rvalues/
atSNP	CPCP	https://github.com/chandlerzuo/atSNP
MBASIC	CPCP	https://github.com/keleslab/mbasic
ENIGMA-Git	ENIGMA	https://github.com/ENIGMA-git/ENIGMA
ENIGMA-DTI	ENIGMA	https://www.nitrc.org/projects/enigma_dti/
ENIGMA-Vis	ENIGMA	http://enigma.usc.edu/research/enigma-vis/
Enigma-Viewer	ENIGMA	http://enigma-viewer.org/About_the_projects.html
ENIGMA-Shape	ENIGMA	http://bit.ly/1RSoUyP
Solar-Eclipse	ENIGMA	http://solar-eclipse-genetics.org/development-team.html
BioGPS	HeartBD2K	http://biogps.org/
Mark2Cure	HeartBD2K	https://mark2cure.org/
MyGene.info	HeartBD2K	https://mygene.info
ProteinInference	HeartBD2K	https://github.com/mpc-bioinformatics/pia
Guten Tag	HeartBD2K	http://fields.scripps.edu/downloadfile2.php?name=GutenTag&filename=GutenTag.zip&id=3
Colander	HeartBD2K	http://www.proteomicswiki.com/wiki/index.php/Colander
DeBunker	HeartBD2K	http://fields.scripps.edu/downloadfile2.php?name=DeBunker&filename=debunker.tar.gz&id=17
GeneWiki	HeartBD2K	https://en.wikipedia.org/wiki/Portal:Gene_Wiki
Charge_Prediction_Machine (CPM)	HeartBD2K	http://fields.scripps.edu/downloadfile2.php?name=Charge_Prediction_Machine&filename=CPM.txt&id=18
Unitemare	HeartBD2K	http://fields.scripps.edu/downloadfile2.php?name=Unitemare&filename=Unitemare.pl&id=19
PatternLab for Proteomics	HeartBD2K	http://www.patternlabforproteomics.org
YADA	HeartBD2K	http://fields.scripps.edu/yada/
MyVariant.info	HeartBD2K	https://myvariant.info
Knowledge.bio	HeartBD2K	http://knowledge.bio
RawExtractor	HeartBD2K	http://fields.scripps.edu/rawconv/
Census	HeartBD2K	http://fields.scripps.edu/census/
ProLuCID	HeartBD2K	http://fields.scripps.edu/downloadfile2.php?name=ProLuCID&filename=&id=12
SEQUEST	HeartBD2K	http://proteomicswiki.com/wiki/index.php/SEQUEST

Name	Center	Link
DTASelect	HeartBD2K	http://fields.scripps.edu/DTASelect/
PINT	HeartBD2K	https://github.com/proteomicsyates/PINT
PSEA-Quant	HeartBD2K	http://pseaquant.scripps.edu
IP2	HeartBD2K	http://www.integratedproteomics.com
BD2KPubMed	HeartBD2K	http://www.heartproteome.org/pubmed/
Proturn	HeartBD2K	http://www.heartproteome.org/proturn/
Branch	HeartBD2K	https://biobranch.org/
OncoRep	HeartBD2K	https://bitbucket.org/sulab/oncorep
Omics Pipe	HeartBD2K	https://bitbucket.org/sulab/omics_pipe
PRIDE	HeartBD2K	http://www.ebi.ac.uk/pride/archive/
ProteomeXchange	HeartBD2K	http://www.proteomexchange.org
Sage Synapse	HeartBD2K	https://www.synapse.org
Gene Expression Atlas	HeartBD2K	https://www.ebi.ac.uk/gxa/home
PSICQUIC	HeartBD2K	http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml
MetaboLights	HeartBD2K	http://www.ebi.ac.uk/metabolights/
Proturn	HeartBD2K	http://www.heartproteome.org/proturn/
Aztec	HeartBD2K	https://aztec.bio
Reactome	HeartBD2K	http://www.reactome.org
COPaKB	HeartBD2K	http://www.heartproteome.org/copa/splash.html
ClusterEng	KnowEnG	http://education.knoweng.org/clustereng/
ClusType	KnowEnG	http://shanzhenren.github.io/ClusType/
SegPhrase	KnowEnG	https://github.com/shangjingbo1226/SegPhrase
AutoPhrase	KnowEnG	https://github.com/shangjingbo1226/AutoPhrase
Saul	KnowEnG	https://github.com/IllinoisCogComp/saul
Zenvisage	KnowEnG	http://zenvisage.github.io/
DataSpread	KnowEnG	http://dataspread.github.io/
Harmonizome	LINCS DCIC	http://amp.pharm.mssm.edu/Harmonizome/
Enrichr	LINCS DCIC	http://amp.pharm.mssm.edu/Enrichr/
GEO2Enrichr	LINCS DCIC	http://amp.pharm.mssm.edu/g2e/
LINCS Data Portal	LINCS DCIC	http://lincsportal.ccs.miami.edu/dcic-portal/
piLINCS	LINCS DCIC	http://eh3.uc.edu/pilincs/#/
LINCS Information Framework (LIFE)	LINCS DCIC	http://life.ccs.miami.edu/life/
iLINCS	LINCS DCIC	http://www.ilincs.org/ilincs/
LINCS Canvas Browser	LINCS DCIC	http://www.maayanlab.net/LINCS/LCB/#.WKY3jhiZMdU
Drug/Cell-line Browser	LINCS DCIC	http://www.maayanlab.net/LINCS/DCB/
Network2Canvas	LINCS DCIC	http://maayanlab.net/N2C/#.WKY5thiZMdU
Docent	LINCS DCIC	http://amp.pharm.mssm.edu/public/docent/
LINCS Project Mobile	LINCS DCIC	http://lincsproject.org/LINCS/mobile
Phosphosite Plus	LINCS DCIC	http://www.phosphosite.org/
Harmonizome Mobile App	LINCS DCIC	http://amp.pharm.mssm.edu/Harmonizome/
L1000CDS2	LINCS DCIC	http://amp.pharm.mssm.edu/L1000CDS2/#/index
L1000CDS2 (API)	LINCS DCIC	http://amp.pharm.mssm.edu/L1000CDS2/#/index
Gen3va	LINCS DCIC	http://amp.pharm.mssm.edu/gen3va/
CREEDS	LINCS DCIC	http://amp.pharm.mssm.edu/CREEDS/
SEP L1000	LINCS DCIC	http://maayanlab.net/SEP-L1000/

Name	Center	Link
Slicr	LINCS DCIC	http://amp.pharm.mssm.edu/Slicr/#/search
PAEA	LINCS DCIC	http://amp.pharm.mssm.edu/PAEA/
GUIdock	LINCS DCIC	https://github.com/WebDataScience/GUIdock
GR Browser	LINCS DCIC	http://grcalculator.org/
Repurposing App	LINCS TG	https://clue.io/repurposing-app
ConnectivityMap	LINCS TG	https://clue.io/cmap
Mmap	MD2K	http://poloclub.gatech.edu/mmap
mCerebrum (MD2K Mobile Software Platform)	MD2K	https://github.com/MD2Korg
OpenSim	Mobilize	http://opensim.stanford.edu
DeepDive	Mobilize	http://deepdive.stanford.edu
SNAP	Mobilize	https://snap.stanford.edu
CVX	Mobilize	http://cvxr.com/cvx/
CVXPY	Mobilize	https://github.com/cvxgrp/cvxpy
DCCP	Mobilize	https://github.com/cvxgrp/dccp
GLRM	Mobilize	https://github.com/madeleineudell/LowRankModels.jl
ECOS	Mobilize	https://github.com/embotech/ecos
SCS	Mobilize	https://github.com/cvxgrp/scs
CVXGEN	Mobilize	http://cvxgen.com/docs/index.html
POGS	Mobilize	https://github.com/foges/pogs
Gamsel	Mobilize	https://cran.r-project.org/web/packages/gamsel/index.html
Glmnet	Mobilize	https://cran.r-project.org/web/packages/glmnet/index.html
softImpute	Mobilize	https://cran.r-project.org/web/packages/softImpute/index.html
Sparsenet	Mobilize	https://cran.r-project.org/web/packages/sparsenet/index.html
SvmPath	Mobilize	http://web.stanford.edu/~hastie/Papers/SVMPATH/
LARS	Mobilize	http://web.stanford.edu/~hastie/Papers/LARS/
gam	Mobilize	https://cran.r-project.org/web/packages/gam/index.html
mda	Mobilize	https://cran.r-project.org/web/packages/mda/index.html
impute	Mobilize	http://bioconductor.org/packages/impute/
PIC-SURE RESTful API	PIC-SURE	https://bd2k-picsure.hms.harvard.edu
ExAC REST API	PIC-SURE	http://exac.hms.harvard.edu
NHANES Dataset Explorer	PIC-SURE	https://nhanes.hms.harvard.edu/transmart/datasetExplorer/index

APPENDIX – II

BD2K Centers of Excellence: NIH IC Relevance

BD2KCCC has requested information from all 13 Centers of Excellence regarding the relevance of their work with respect to NIH ICs. Provided below is a list of one or more NIH ICs and disease foci that each BD2K Center of Excellence is related to. One could appreciate that many centers cover a multitude of scientific themes that support the mission of multiple ICs.

1. U54 EB020406: Big Data for Discovery Science (BDDS)
PD: Arthur W. Toga, University of Southern California

Relevant IC:
Our center, entitled Big Data for Discovery Science, will lead a new paradigm for interacting with large biomedical data types and scales – from 'omes' to 'organs'. We see our efforts as relevant to multiple ICs. Examples are given below:

- NIGMS – Development of Trans-Proteomic Pipeline for proteomics and proteogenomics data analysis; integrated multi-omic models of cells; pilot for creating reproducible workflows using Docker containers for NIH Commons.
- NINR – Study of Complexity and Self-Management of Chronic Disease (CSCD)
- NINDS – Cholinergic Mechanisms of Gait Dysfunction in Parkinson's Disease; identifying a diffusion MRI signature of tau aggregation in tauopathies and the effects of tau aggregation on hippocampal connectivity; neuroimaging analytics for traumatic brain injury (TBI), see http://tinyurl.com/hgzkaoe.
- NIA – Neurodegenerative Diseases and Dementia; TRENA for identifying drug targets in Alzheimer's disease.
- NCI – (PBQ4) Single cell analysis strategy for monitoring drug responses of tumor cells; Sex Differences Supplement: (PBQ4) Single Cell Analysis Strategy for Monitoring Drug Responses of Tumors.
- NCATS – Biomedical data translator technical feasibility assessment and architecture design.
- NHGRI – Large-scale, cloud-based genomics analysis.
- NICHD – Comparing how using a phonologic strategy vs an orthographic strategy to perform a word recognition task differentially makes use of the existing structural brain networks in three populations: children with dyslexia, their age-matched controls, and their reading-level matched controls; participating as a partner in the Autism Centers of Excellence program.
  see http://journal.frontiersin.org/article/10.3389/fpsyt.2016.00205/full.
- NIDDK – (rebuildingakidney.org, gudmap.org).
- NIDCR – (facebase.org).

2. U54 HG008540: Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data.
PD: Gregory F. Cooper, University of Pittsburgh School of Medicine
MPI: Ivet Bahar, University of Pittsburgh School of Medicine

The CCD causal discovery tools are applicable to the analysis of big data in all the NIH ICs. The specific biomedical projects being investigated in the CCD (cancer, lung, and brain fMRI projects) are applicable to specific ICs, as indicated in the list below. The ICs to which new algorithm development is applicable are also listed.

- NHLBI – lung project.
- NCI – cancer project, lung project.
- NINDS – brain fMRI project.
- NIBIB – brain fMRI project, cancer project, lung project, algorithm development.
- NIA – brain fMRI project, lung project.
- NIGMS – cancer project, lung project, algorithm development.
- NICHD – brain fMRI project.
- NIMH – brain fMRI project.
- NIEHS – lung project.
- NHGRI – cancer project, lung project.
- NIDDK – lung project (fibrosis in other organs).
- NCATS – lung project.
- NLM – algorithm development.

3. U54 AI117925: Center for Expanded Data Annotation and Retrieval (CEDAR)
PD: Mark A. Musen, Stanford University School of Medicine

CEDAR's work to ease the burden of metadata authoring is relevant to every IC at the NIH. Every IC is concerned with investigation to create new experimental data sets, and every IC has a commitment to open science and to the "FAIR" exchange of scientific information. Although much of CEDAR's initial activity has been driven by needs in immunology and, thus, has been within the purview of NIAID, our approach is not tied to any particular disease focus. As the NLM assumes responsibility for data science research at the NIH, aspects of CEDAR's activity will support NLM's mission in this area.

4. U54 AI117924: Center for Predictive Computational Phenotyping (CPCP)
PD: Mark W. Craven, University of Wisconsin – Madison

Relevance to ICs:
Research projects in CPCP have a number of disease/health foci including:

- NCI – risk stratification for improved, targeted detection and diagnosis of breast cancer.
- NIAID – characterizing preclinical brain changes leading to Alzheimer's disease.
- NIGMS – identifying drug repositioning candidates using electronic health records.
- NHLBI – predicting adverse events from electronic health records including VTEs and asthma exacerbations.
- NHGRI – elucidating aspects of genome architecture including regulatory SNP detection and long-range regulatory interactions.
- NIDDK – understanding hematopoietic stem cell genesis.
- NIAID – multidimensional characterization of virus replication.

5. U54 EB020403: ENIGMA Center for Worldwide Medicine, Imaging, and Genomics
PD: Paul M. Thompson, The University of Southern California School of Medicine.

Relevance to ICs/disease focus:
ENIGMA's 33 Working Groups target themes of interest to:
- NIMH – Depression (MDD), Bipolar disorder, Schizophrenia, ADHD, Anorexia Nervosa, Obsessive Compulsive Disorder, Tourette syndrome, Anxiety (Panic Disorder, Seasonal Affective Disorder, Generalized Anxiety Disorder), Irritability, PTSD (posttraumatic stress disorder), Aggression, Cross Disorders Group 1.
- NIBIB – imaging genetics, DTI (diffusion tensor imaging), resting-state fMRI, EEG, hippocampal subfield algorithms, shape analysis.
- NICHD – 22q deletion syndrome, ADHD, autism spectrum, pediatric OCD, Tourette's syndrome, early onset-psychosis, lifespan, Lysosomal Storage Diseases (LSDs), early life trauma, Cross Disorders Group 2
- NIA – FTD (fronto-temporal dementia), Parkinson's Disease, GWAS of aging (Plasticity).
- NINDS – Epilepsy, HIV/AIDS, Ataxia, Stroke Recovery, TBI (traumatic brain injury).
- NIDA/NIAAA – Addiction (including alcohol, nicotine, cocaine, meth), Laterality.
- NIAID – HIV/AIDS.
- NHGRI – GWAS, Epigenetics, 22q deletion syndrome, CNV (copy number variation), Evolution, Lysosomal Storage Diseases (LSDs).
- FIC – Worldwide 35-country Collaborations.
- NLM – big data in medicine.

6. U54 GM114833: A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform.
PD: Peipei Ping, UCLA School of Medicine
MPI: Andrew Su, The Scripps Research Institute
Merry Lindsey, University of Mississippi Medical Center
Karol Watson, UCLA School of Medicine

Relevance to ICs/disease focus:
- NIGMS – given our projects on advancing the understanding of biological processes in cardiovascular diseases, rare diseases, and in aging; as well as our major efforts in enhancing approaches in disease diagnosis, treatment, and prevention.

- NHLBI – given the relevance of projects with use cases focusing on cardiovascular disease (CVD).
- NIA – given one of our use case scenarios is an aging cohort; as well as our study focus on aging-related conditions related to CVD.
- NIAMS – given our study on physical activities and heart failure; as well as our collaborative work on lung failure and cardiomyopathy in muscular dystrophy.
- NIDDK – given our study focus on mitochondrion, and the role of mitochondrial proteins in metabolic syndrome and diabetes settings of CVD.
- NIEHS – given the evolving relationship of environmental factors as major risk factors to impact the outcome of CVD development and treatment.
- NLM – given our investigative focus on text mining and indexing-related work being developed; and our efforts in enabling datasets/toolsets to be in compliance with FAIR Principles.

7. U54 GM114838: KnowEnG, a Scalable Knowledge Engine for Large-Scale Genomic Data.
PD: Jiawei Han, Univ. of Illinois at Urbana-Champaign
Co-PD: Saurabh Sinha, Univ. of Illinois at Urbana-Champaign Institute for Genomic Biology
MPI: Jun S. Song, Univ. of Illinois at Urbana-Champaign Institute for Genomic Biology
Richard M. Weinshilboum, Mayo Clinic.

Our center's research on cancer pharmacogenomics is most relevant to the NCI. Also, our collaborative research with the HeartBD2K center is relevant to the NHLBI.

8. U54 HL127624: Data Coordination and Integration Center for BD2K-LINCS (BD2K-LINCS DCIC).
PD: Avi Ma'ayan, Icahn School of Medicine at Mount Sinai
MPI: Mario Medvedovic, University of Cincinnati Medical Center
Stephan C. Schurer, University of Miami & The Scripps Research Institute

- NHLBI – The Library of Integrated Network-Based Cellular Signatures (LINCS) project is expected to produce masses of data collected from human cells and tissues perturbed with drugs and other molecules. This effort is relevant to multiple ICs including NHLBI.

9. U54 EB020404: Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K).
PD: Santosh Kumar, University of Memphis

Below is a list of health outcomes and the corresponding institute that is being addressed by MD2K:

- NHLBI – Congestive heart failure.
- NCI, NIMH, NIDA – Smoking.
- NIDA – Cocaine Use.
- NIDCR – Oral Health.
- NIDDK – Overeating.
- NIBIB – Health technology.

10. U54 EB020405: Mobilize Center.
PD: Scott L. Delp, Stanford University

- NIAMS, NCI, NIMH – Monitoring physical activity and other health related behaviors in free-living populations using commercial smartphones and activity monitors.
- NIA, NIMHD – Designing interventions to increase activity and reduce sedentary behavior in older adults and in minority and other under-served populations.
- NIEHS – Identifying environmental factors that impact physical activity levels based on mobile health technologies and designing effective interventions to increase activity based on these factors.
- NIAMS, NINDS, NIA – Developing new approaches to monitor and improve the treatment of osteoarthritis, stroke, and other age-related disabilities that impair mobility.
- NICHD – Developing new techniques to automatically monitor and classify activity, sleep, and sedentary time in children and applying to test the efficacy of interventions to increase activity in children.
- NICHD – Improving the diagnosis and treatment of movement disorders in children with cerebral palsy.
- NIBIB, NIDDK – Improving the diagnosis of diabetes through the integration of data from commercial activity monitors and omics data.

- NIMH – Discovering effective counseling tactics through the analysis of text-based crisis help lines.
- NIBIB, NIAMS – Developing data science tools to automatically analyze imaging data for diagnosis of OA.
- NIBIB – Creating tools to simulate the musculoskeletal system and integrate data from a wide variety of sensors.
- NLM, NIGMS – Developing tools to automatically extract information from unstructured data in clinical notes and biomedical literature.
- NLM – Creating web infrastructure to build online communities to share biocomputational data and tools.

11. U54 HG007990: Center for Big Data in Translational Genomics
PD: David H. Haussler, The University of California Santa Cruz
MPI: David Patterson, The University of California Santa Cruz
Laura J. van 't Veer, The University of California Santa Cruz

Our Center's disease focus has application to NCI and NHLBI (TOPMed).

12. U54 HG007963: Patient-Centered Information Commons: Standardized Unification of Research Elements (PIC-SURE)
PD: Isaac Kohane, Harvard University Medical School

Funded by the NIH Big Data to Knowledge (BD2K) Program, we propose to create a massively scalable toolkit to enable large, multi-center Patient-centered Information Commons (PIC) at the local, regional, and national scales, where the focus is the alignment of all available biomedical data per individual. Such a Commons is a prerequisite for conducting the large-N, Big Data, longitudinal studies essential for understanding causation in the Precision Medicine framework while simultaneously addressing key complexities of the Patient-Centric Outcome Research studies required under Affordable Care Act (ACA). Our proposal is solidly grounded in our experience over the last 25 years of harnessing clinical care data to the research enterprise.

- NHGRI
- NCATS
- NEI
- NIAID
- NIDDK
- NIEHS

13. U54 HL127366: Broad Institute LINCS Center for Transcriptomics (LINCScloud)
PD: Todd Golub, Broad Institute of MIT & Harvard
MPI: Aravind Subramanian, Broad Institute of MIT & Harvard

The overarching goal of the Broad Institute’s transcriptomics center is the development of comprehensive signatures of cellular states that can be used by the entire research community to understand protein function, small-molecule action, physiological states, and disease states. Our center will create the world’s most comprehensive resource of perturbational signatures. This will include 1.4 million L1000 genetic (CRISPR knock-out, shRNA knock-down and ORF overexpression) and small-molecule (drug and tool compound) perturbations spanning 50 cell types of varied lineage. We will make it possible for biologists and computational scientists worldwide to interact with the data by creating user-friendly apps that are designed to facilitate biological discovery. We are collaborating with researchers at the Environmental Protection Agency (EPA) and the National Institute of Environmental Health Sciences (NIEHS) to apply transcriptional signatures and associated metadata to inform the safety of compounds in the environment. These efforts have produced rich annotations for over 400 small molecules actively being studied by the toxicology community.

- NHGRI
- NIEHS