Big Data — view from NIA/NIH

Nina Silverberg

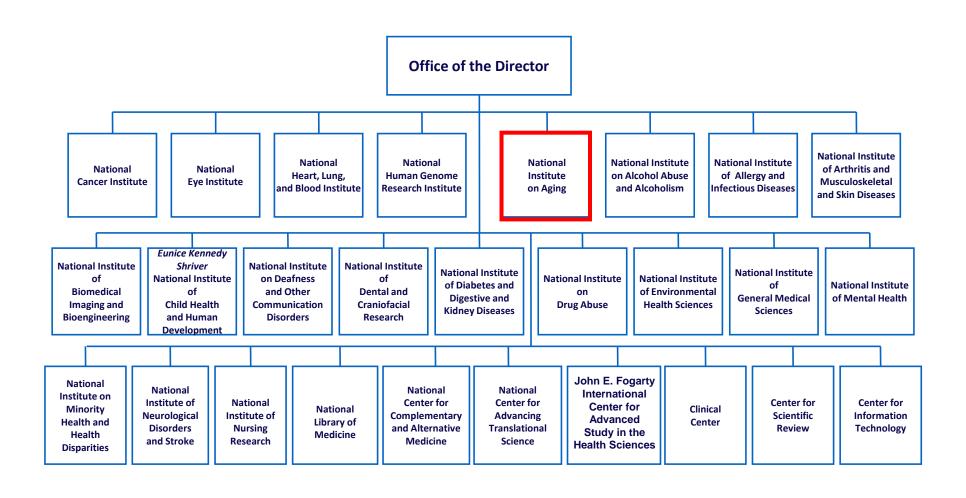
Program Director, Alzheimer's Disease Centers Program

Division of Neuroscience, NIA

May 19

The Center for Network and Storage Enabled Collaborative Computational Science Symposium

National Institutes of Health



27 Institutes and Centers (ICs)

NIH Mission

To seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.

FY2017 Priorities:

- Foundation for Discoveries: Basic Research
- The Promise of Precision Medicine
- Applying Big Data and Technology to Improve Health
- Stewardship to Inspire Public Trust



Initiatives that Support Computational and Mathematical Sciences

Biomedical Information Science and Technology Initiative (BISTI)

Promote the optimal use of computer science and technology to address problems in biology and medicine by fostering collaborations and interdisciplinary initiatives (bisti.nih.gov)

Big Data to Knowledge Initiative (BD2K)

Develop new approaches, standards, methods, tools, software and competencies that will enhance the use of biomedical Big Data by supporting research, implementation and training in the data sciences (datascience.nih.gov/bd2k)

Interagency Modeling and Analysis Group (IMAG)

Provide an open forum for communication among government representatives for transagency activities that have a broad impact in science (imagwiki.nibib.nih.gov)

NSF/NIH Joint program in Mathematical Biology

Bring mathematics and statistics into the core of biological and biomedical research and to broaden the use of innovative mathematics in understanding life processes.



FY17 Priorities for NIH

Big Data to Knowledge (BD2K)

- Coordinate access to and analysis of the many types of biological and behavioral 'big data' being generated by biomedical scientists
- Develop innovative and transformative computational approaches, tools, and infrastructures to make 'big data' and data science a prominent component of biomedical research
- Enable data sharing and utilization through the development of a new shared, interoperable cloud computing environment: the 'Commons'

Big Data to Knowledge (BD2K)

- launched in 2014 to:
 - facilitate broad use of biomedical big data
 - develop and disseminate analysis methods and software
 - enhance training relevant for large-scale data analysis
 - and establish centers of excellence for biomedical big data.
 - supported initial efforts toward making data sets "FAIR"
 - Findable, Accessible, Interoperable, and Reusable.
- Second phase: will test the feasibility of, and develop best practices for, making NIH funded datasets and computational tools available in a shared space that multiple scientists can access remotely.



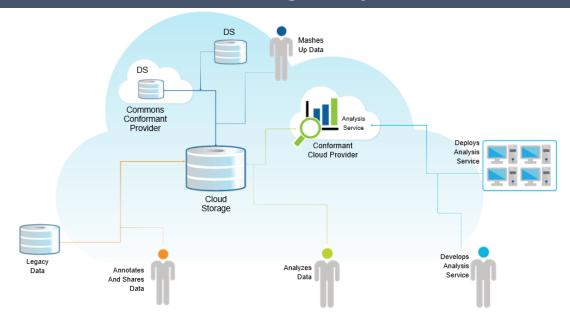
Role of Data Sciences at NIH

Enable broad data sharing and reuse of data Findable, Accessible, **FAIR** Interoperable, and Re-usable (FAIR) Support biomedical discovery by enabling the sharing of digital Commons objects Enable an effective and diverse biomedical, data science workforce **Training** Develop An NIH Vision For Economic, Technical, And Social Sustainability Stewardship Of Biomedical Data Repositories.



The Commons as Innovation Accelerator

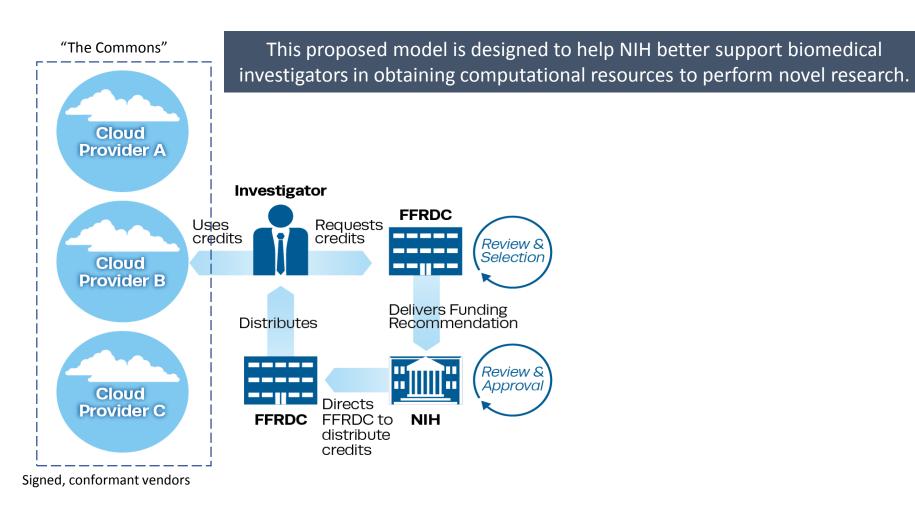
NIH proposes a **community-owned** cloud-based electronic ecosystem ("**Commons**") where researchers can **store**, **share**, **and utilize** their own, and others', sharable Digital Objects.



How can this best be **supported** so as to reduce long-term costs, increase re-use of Digital Objects, and promote the overall scientific output of the nation?



Commons Credits Pilot





It's Easy to Participate in the Commons Credits Pilot

Each step is designed to be as simple and low effort as possible to help reduce the barriers to entry and participation.

- Applications are < 2 pages, and should be easy to complete in less than 1 day.
- Credits will be disseminated within 8 weeks post-cycle close, and are available for 12 months.
- Investigators focus on the science, MITRE handles invoicing.
- Participants' forum on Portal for sharing and discussion.

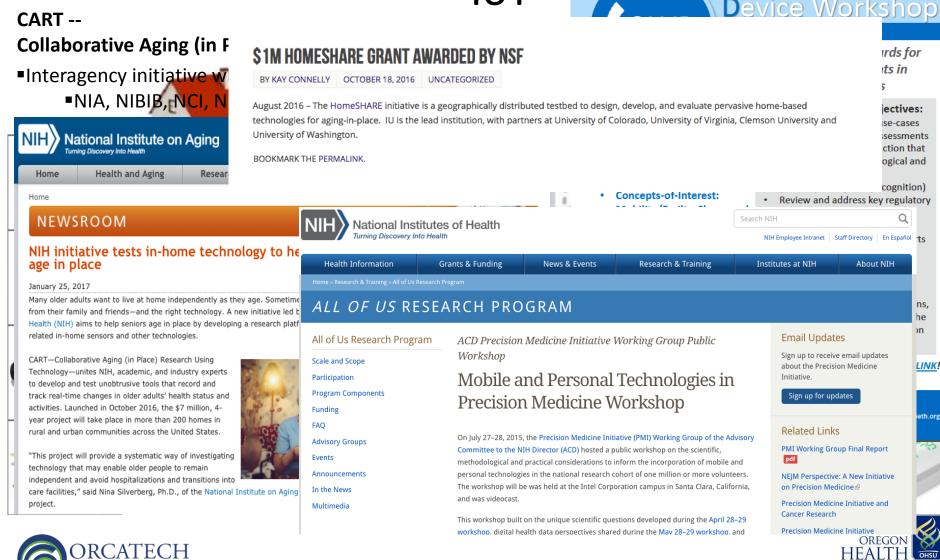
commons credits@mitre.org

Ongoing research which utilizes big data storage: Examples

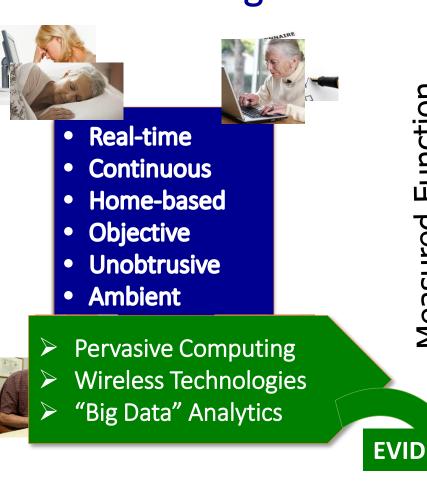
High Interest in Digital Technologies

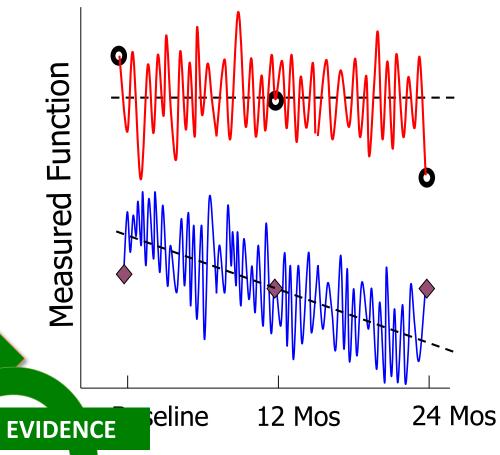
IoT





Which has brought us to BMDs in Trials...

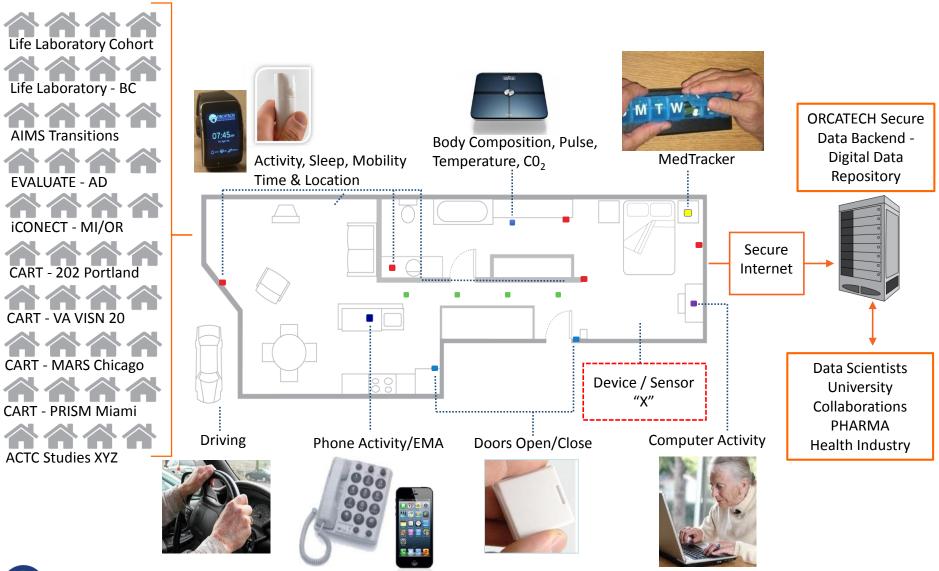








Technology 'agnostic' pervasive computing platform for continuous home-based assessment and Tx





Sample project (two sites, one in MI)

- R01 project,
- using skype-like video chats to improve cognitive function of socially isolated elderly.
- Each conversational session audio/video recorded for later analyses
- 30 minute video recording in mp4 format requires 300 megabyte of storage. 4 times per week = 1200 Mega byte of storage per week. 1200 X 180 subjects X 24 weeks require 5.2 terabyte of storage space (just for the 1st 6 months)

Video Chat



Systems and Data Storage

- Conversational sessions will be recorded for audio/visual analyses (mp4)
- 30 minutes of video chat = 300 megabyte data
- 300 megabyte X 4 times /week X 24 months X 180 subjects = 5.2 terabyte of data! (for the 1st 6 months)
- Currently exploring storage options

NIAGADS GenomicsDB

SEARCH

Search for a gene, variant, or region

Examples - Gene: APOE - Variant: rs6656401 - Region: chr19:45309039-45512650

The NIAGADS GenomicsDB is a searchable annotation resource that provides access to publicly available NIAGADS summary statistics datasets for Alzheimer's Disease and related neuropathologies. These data are curated along with variant and gene annotations and AD-revelant functional genomics datasets, allowing AD researchers to easily identify and interpret interesting genomic regions via interactive search strategies and the NIAGADS genome browser.

What would you like to do? What's New? Explore the region around a gene or SNP on the genome browser. e.g., APOE WIEW WEW Perform pathway or functional enrichment analysis on a list of genes. Genes.





generates 3D MRI volumes. The software comes pre-trained on a provided data set but can be retrained to work with your desired regions of interest.



and manipulate geometric shapes and associated data values. It simplifies the creation of application programs by providing a readymade set of support routines.

FEATURED STUDIES

The IDA contains data on 35,589 subjects from 106 studies









































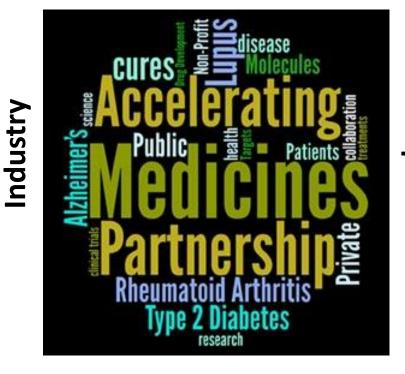
SYSTEMS APPROACHES FOR TARGET DISCOVERY AND VALIDATION

Suzana Petanceska PhD

Accelerating Medicines Partnership Alzheimer's Disease Program

https://www.nia.nih.gov/alzheimers/amp-ad

Government



Non-profit

AMP-AD Partners





















ALZHEIMER'S DISEASE - Target Discovery and Preclinical Validation Project

- ☐ Discover and carry out preclinical validation of novel disease-relevant therapeutic targets by integrating the analyses of large-scale molecular data from human brain/blood samples with network modeling approaches and experimental validation.
- ☐ Enable rapid and broad sharing of data.

ALZHEIMER'S DISEASE - Target Discovery and Preclinical Validation Project

☐ The project is a consortium of 6 multi-institutional, multidisciplinary research teams supported by NIA grants.

- □ The teams are applying cutting-edge systems and network biology approaches to integrate multidimensional human "omic" data (genomic, proteomic, metabolomic) from ~2,500 human brains/~1000 blood samples from all stages of the disease with clinical and pathological data to:
 - discover and select novel therapeutic targets for Alzheimer's disease
 - gain a systems-level understanding of the gene, protein, and metabolic networks within which these targets operate
 - evaluate their druggability in cell-based and animal models

ALZHEIMER'S DISEASE - Target Discovery and Preclinical Validation Project

Generate

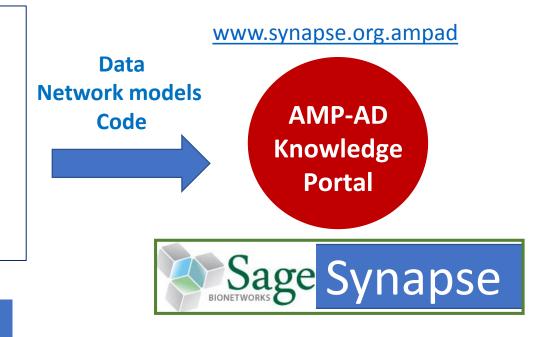
High-dimensional multi-omic data: ~2,500 human brains;~1000 blood samples

Integrate

Molecular profiling
Predictive Modeling
Experimental validation

6 Academic Teams

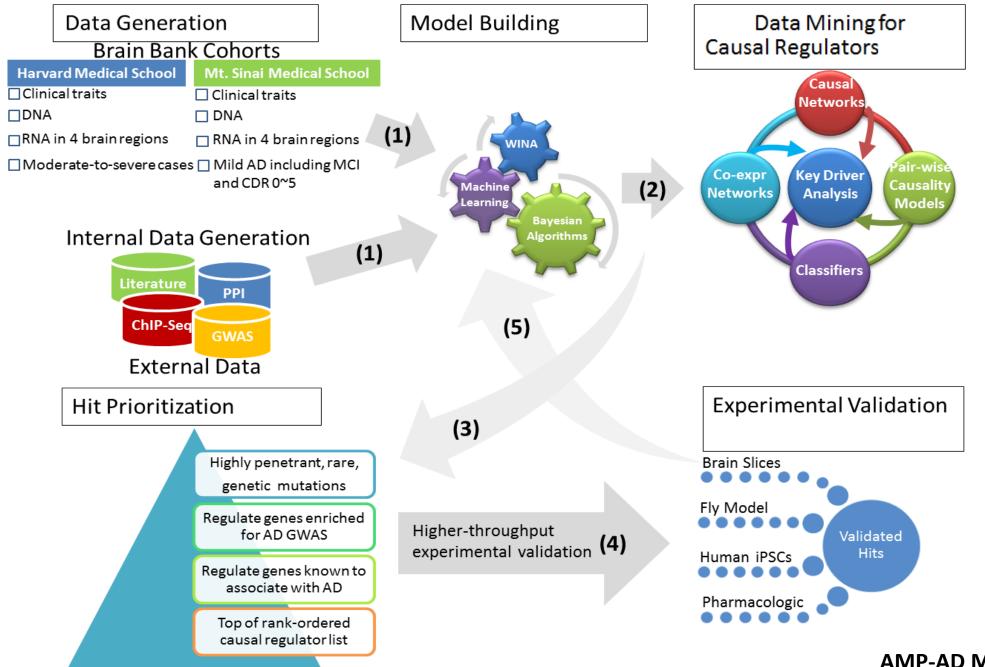
- NIA U01/R01grants -



ALZHEIMER'S DISEASE - Target Discovery and Preclinical Validation Project

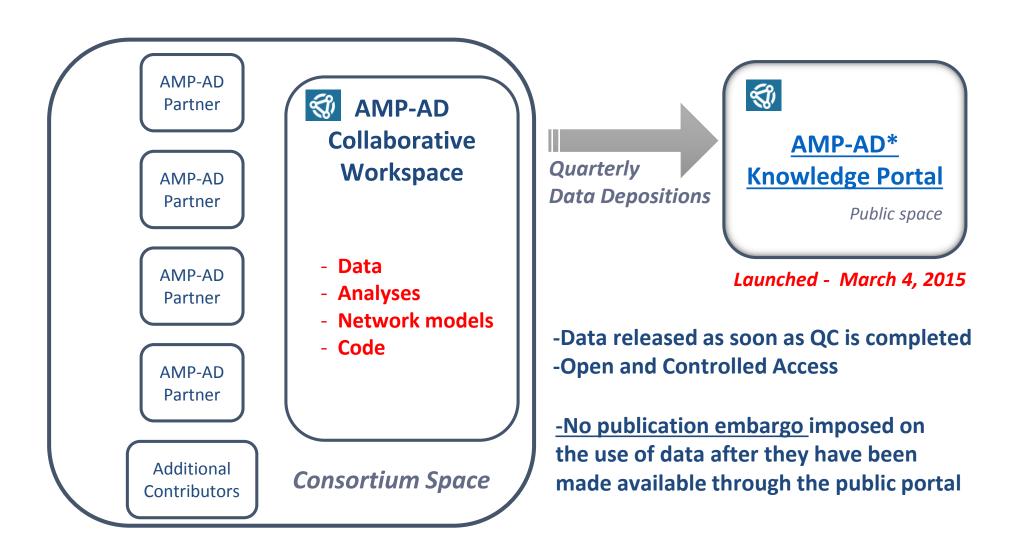
Academic Teams	Broad- Rush	Mt Sinai	UFL/ISB /Mayo	Emory	Duke	Harvard/ MIT
Principal Investigators	De Jager, Bennett	Schadt, Zhang	Golde, Price, Taner	Levey	Kaddurah- Daouk	Yankner, Tsai
Human Data source	ROSMAP	Mt Sinai Brain Bank	Mayo Brain Bank	All	ADNI	ROSMAP
Molecular Data Types	RNAseq	RNAseq Whole exome seq	RNAseq	All Proteomics	Metabolomic	Txpn Factors
Target Identification	Bayesian networks	Bayesian networks	Innate Immunity Networks	Bayesian Networks	Systems analysis	REST
Preclinical Validation	iPSCs Cell lines	iPSC, drosophila, mouse	mouse	Mouse, cell culture, drosophila	NA	mouse

Data Enablement and Coordination of Collaborative Analyses: Sage Bionetworks, Principal Investigator – Lara Mangravite



AMP-AD Mt.Sinai team: Project Workflow







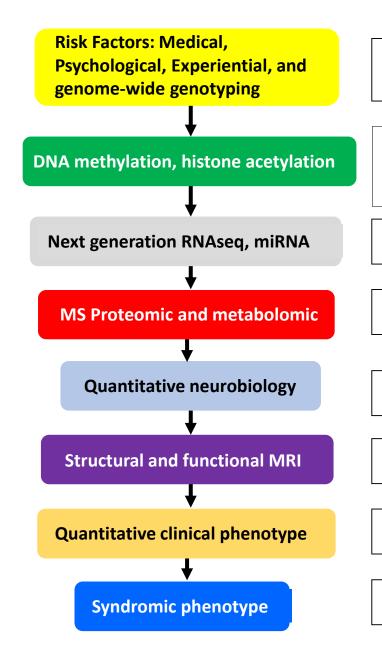
AMP-AD Knowledge Portal

	HUMAN TISSUE	Diagnosis	Assay
	Prefrontal Cortex	 Alzheimer's Disease Mild Cognitive Impairment Parkinson's Disease Amyothrophic Lateral Sclerosis Corticobasal Degeneration Frontotemporal Dementia Dementia with Lewy Bodies 	 RNAseq Gene Expression array miRNA array ChIPseq DNA Methylation array Proteomics Confocal Imaging SNP genotypes Proteomics Whole Exome Seq
	Visual Cortex	Alzheimer's Disease	Gene Expression ArraySNP genotypes
	Temporal Cortex	Alzheimer's DiseaseProgressive Supranuclear PalsyParkinson's Disease	RNAseqSNP genotypes
	Cerebellum	Alzheimer's DiseaseProgressive Supranuclear PalsyParkinson's Disease	• RNAseq
	Superiour Temporal Gyrus	Alzheimer's Disease	RNAseqWhole Exome Seq
Pa	arahippocampal Gyrus	Alzheimer's Disease	• RNAseq
	Serum	Alzheimer's DiseaseMild Cognitive Impairment	Metabolomics

MODELS	Туре	Assay
Drosophila	TAUTYROBPTREM2	ProteomicsRNAseq
Mouse	APPTAUPSEN1HDAC1db/dbStroke	RNAseqGene Expression arrayBehavioral analysisElectrophysiology
CELLULAR MODELS - Human - Rat	 iPSC iPSC derived organoids Microglia Macrophages Astrocytes APP/PSEN1, SNRNP70 transduced neurons SH-SY5Y 	 RNAseq ChIPseq ELISA (Aβ) Proteomics

Religious Orders Study and Rush Memory and Aging Project

- Two cohort studies of aging and AD ongoing for 20+ years
- >3,000 older persons without [known] dementia from across the USA
- All agreed to annual detailed clinical evaluation for common chronic conditions of aging with detailed evaluation of risk factors, and blood donation
- All agreed to organ donation at death
- > 900 cases incident MCI
- > 700 cases incident AD dementia
- > 1,200 autopsies



Genotypes

Affy 6.0/Illumina Quad 1000G imputation

Genotypes

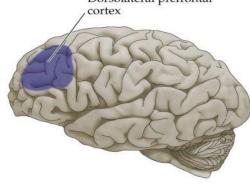
Whole Exome Sequencing

DNA Methylation Illumina 450K Histone H3K9Ac

ChIP-Seq

RNA profile mIRNA & RNAseq

LC/MS profile Lipids, proteins Dorsolateral prefrontal



Neuropathology

AD, CVD, LBD, HS, TDP

Resilience Markers Synaptic proteins, LC

Flair, MP Rage, DTI, SWI, rsfMRI

post-mortem MRI DTI, MP Rage, T2

Cognitive Function 19 tests annually

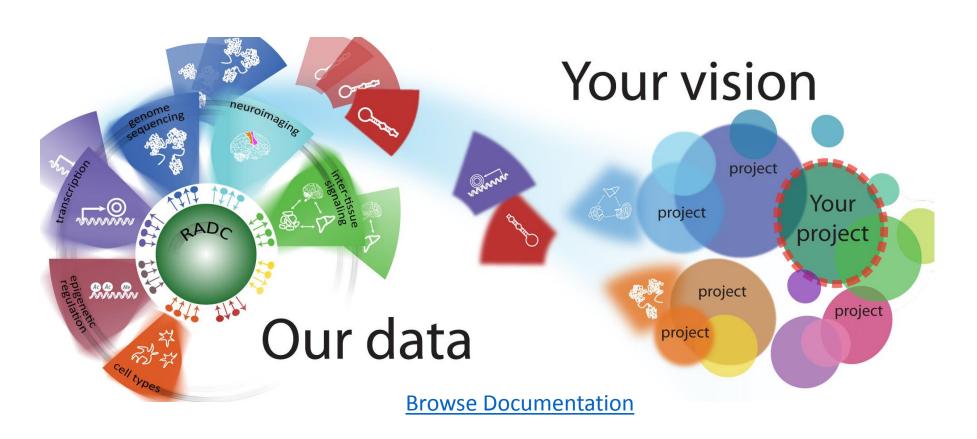
Motor Function Disability BMI, Actigraphy

Dynaport - Gait, Sleep, circadian rhythms, Behavioral Economics, Olfaction,

Clinical Diagnoses AD, Stroke, PD,

RADC Research Resource Sharing Hub

https://www.radc.rush.edu

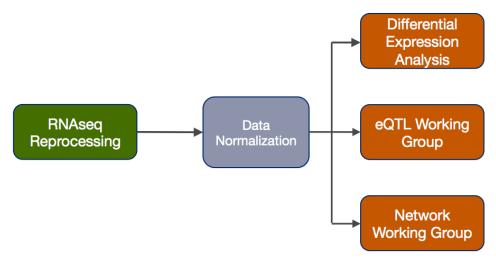


Query Frequency Reports

Request Data/Specimens

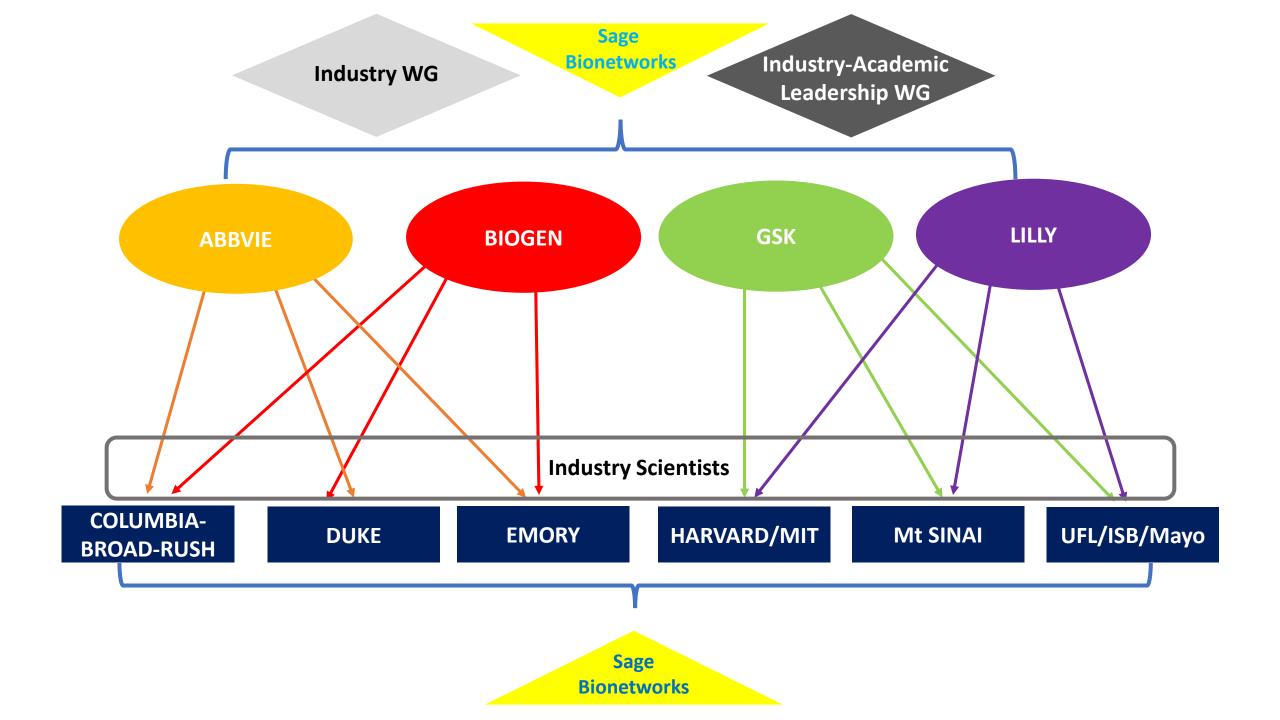
AMP-AD RNASeq Reprocessing WG: Goals and Deliverables

- Enable joint analysis through uniform reprocessing to reduce technical variation across Human RNAseq datasets
- Meta-analysis to inform internal AMP-AD projects and support target selection processes
- Development of a standardized resource for external users



RNAseq reprocessing working group

29 members representing 5 AMPAD academic teams and all 4 industry partners Contacts: <u>kristen.dang@sagebase.org</u> & <u>thanneer.perumal@sagebase.org</u>



NIH's All of Us Initiative





All of Us > About > Program Components

Program Components

Through a set of funding awards, NIH has established the essential components of the All of Us Research Program to build a research cohort of one million or more U.S. volunteers to advance precision medicine. Click the images below to learn more about each of these



Data & Research Center



Participant Center



Health Care Provider Organizations

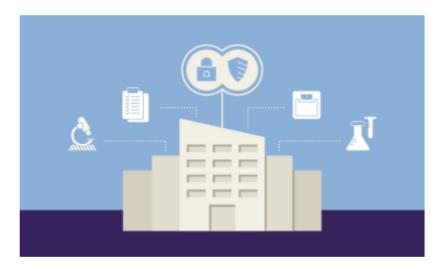


Biobank



Participant Technology Systems Center

Data & Research Center



The Data and Research Support Center will acquire, organize, and provide secure access to what will be one of the world's largest and most diverse datasets for precision medicine research. The center will also provide research support for the scientific data and analysis tools for The Program, helping to build a vibrant community of researchers from community colleges to top health care research institutions and industries, as well as citizen scientists, who can propose studies using this information.

Awardees:

 Vanderbilt University Medical Center, Nashville, Tennessee, working with the Broad Institute, Cambridge, Massachusetts, and Verily, Mountain View, California

Sub-awardees:

- Columbia University Medical Center, New York City
- · Northwestern University Feinberg School of Medicine, Chicago, Illinois
- University of Michigan School of Public Health, Ann Arbor, Michigan
- University of Texas Health Science Center at Houston School of Biomedical Informatics, Houston, Texas



Data Science Home / Funded Programs

Introduction

other data science-related areas.







Introduction

BD2K Centers

Resource Indexing

Enhancing Training

Targeted Software Development

BD2K Centers

The BD2K Centers of Excellence program has established 11 Centers of Excellence for Big Data Computing and one Center that is a collaborative project with the NIH Common Fund LINCS program, the LINCS-BD2K Perturbation Data Coordination and Integration Center, The centers are located all across the United States. They are large-scale projects aiming to develop new approaches, methods, software tools, and related resources. The Centers will also provide training to advance Big Data science in the context of biomedical research. The 12 BD2K Centers function with the other BD2K grantees as a consortium and collaborate with one another, with the purpose of furthering every aspect of the field of biomedical data science research.

BD2K funds research and training activities that support the use of Big Data to advance biomedical research and

discovery. This includes efforts in enhancing training, resource indexing, methods and tools development, and

Read more about the BD2K Centers

Resource Indexing

To harness the full potential of Big Data scientists must be able to readily find, cite, and access existing data and other digital objects, such as software. There is no existing infrastructure or incentive that enables this. These basic goals maximize data use, enable sharing, limit duplication of effort, and allow areas of sparse research coverage to be more readily identified. To advance the infrastructure and policies needed to meet these goals, awards in this area address the challenges of resource discovery, citation, and access.

Read more about BD2K awards supporting Resource Indexing





BD2K IS IN THE iew BD2K News



Funding Opportunities



NIGMS Institutional Predoctoral Training Programs

- Behavioral-Biomedical Sciences Interface
- Bioinformatics and Computational Biology
- Biostatistics
- Biotechnology
- Cellular, Biochemical, and Molecular Sciences
- Chemistry-Biology Interface
- Genetics
- Medical Scientist Training Program (M.D.-Ph.D.)
- Molecular Biophysics
- Molecular Medicine
- Pharmacological Sciences
- Systems and Integrative Biology



Ruth L. Kirschstein National Research Service Award (NRSA)

- Awards honor Dr. Ruth L. Kirschstein, former Director of the National Institute for General Medical Sciences. Aside from Dr. Kirschsteins scientific accomplishments in polio vaccine development, she was a champion of research training and a strong advocate for the inclusion of underrepresented individuals in the scientific workforce
- Individual Predoctoral MD/PhD or Other Dual-Doctoral Degree Fellowship PA-14-150
- Individual Predoctoral Fellowship (PA-14-147)
- Individual Predoctoral Fellowship to Promote Diversity in Health-Related Research (PA-14-148)
- Individual Senior Fellowship (PA-14-151)

PAR 17-032

Translational Bioinformatics Approaches to Advance Drug Repositioning and Combination Therapy Development for Alzheimer's Disease (RO1)

Award Budget - Annual direct costs are capped at \$500K.

Award Project Period - The maximum project period is 5 years.

PROGRAMMATIC GOAL: Establish new research programs that will promote the use of systems-based, data-driven approaches to create a knowledge base needed for successful drug repositioning and combination therapy development for AD.

To this end this funding opportunity announcement is soliciting projects that use of existing or develop novel computational approaches to identify drugs or drug combinations currently used for other conditions with potential to be efficacious in AD and AD-related dementias.

This initiative encourages cross-disciplinary, team-science approach and academia-industry collaborations.

Research scope/Examples of responsive applications

- ➤ Purely computational research aimed at using existing methodology to analyze various types of molecular and clinical data to identify individual drugs or drug combinations with favorable efficacy and toxicity profiles as candidates for repositioning against AD or AD-related dementias.
- ➤ Studies proposing the use of translational bioinformatics approaches to integrate existing data with newly generated molecular data collected from biosamples from legacy trials for AD that have tested the efficacy of repurposed drugs (statins, NSAIDs etc.) for the purpose of identifying the molecular determinants of responder phenotypes.

Research scope/Examples of responsive applications

- ➤ Research that combines computational and experimental approaches to generate data-driven predictions on the efficacy of repurposed drugs or drug combinations, followed by efficacy testing in proof-of-principle animal studies or in proof-of-principle human trials.
- ➤ Research that combines computational and experimental approaches to identify quantitative methods that can assess synergy/additivity of candidate therapeutics, including synergy between drugs and non-pharmacological perturbations.
- ➤ Of particular interest are projects that leverage the network concept of drug targets and the power of phenotypic screening to advance rational drug repurposing and data-driven development of drug combinations based on the ability of single or multiple therapeutic agents to perturb entire molecular networks away from disease states in cell-based and/or animal models.

I The development and testing of new therapeutic agents is outside the scope of this
funding opportunity.

Applicants are expected to adhere to NIH guidelines for rigorous study design and transparent reporting to maximize the reproducibility and translatability of their findings.