Data, Data, Everywhere
AMP Portals and Interoperability

Accelerating Medicines Partnership® Program Symposium
February 6, 2024
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)  
AMP AD

Noel Burtt (Broad)  
AMP CMD

Sweta Ladwa (NHLBI)  
AMP HF

Matt Bookman (Verily)  
AMP PD

Anna Greenwood (Sage)  
AMP data landscape
AMP SBI goals and challenges
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)  
AMP AD

Noel Burtt (Broad)  
AMP CMD

Sweta Ladwa (NHLBI)  
AMP HF

Matt Bookman (Verily)  
AMP PD

Anna Greenwood (Sage)  
AMP data landscape  
AMP SBI goals and challenges
AMP AD
Translational Impact of Open Science and FAIR Data Practices

Suzana Petanceska, PhD
National Institute on Aging

Accelerating Medicines Partnership® Program Symposium
Bethesda, February 5-6, 2024
Enhancing the Reach and Impact of Data in the AD Knowledge Portal

- Commitment to FAIR Principles
- Streamlined Data Governance
- Data deposition post QC
- Data made available to AMP-AD Partner organizations and research community at large at the same time
- No publication embargo
- User-Centered Design
- Community Outreach and Education
AD Knowledge Portal: User Support

Documentation And Help Site

Discussion Forum

Monthly Data Release Notes and Newsletter

Hands On Workshops
NIA Translational Centers Leveraging AMP AD Deliverables

Established in 2016

MODEL-AD

- Developing new, knock-in, mouse models based on GWAS for late-onset AD (LOAD)
- Using AMP AD data and network modules for evaluating the molecular homology with human LOAD
- Preclinical efficacy testing pipeline

https://model-ad.org

Established in 2019

TREAT-AD

- Using AMP AD data and target nominations to develop open-source research tools for studying disease biology and for advancing novel targets into drug discovery
- Open data and distribution of Target Enabling Packages free of IP barriers

https://treatad.org
MODEL-AD

The Model Organism Development and Evaluation for Late-onset Alzheimer’s Disease (MODEL-AD) Consortium was established to maximize human datasets to identify putative variants, genes, and biomarkers for AD; to generate, characterize, and validate the next generation of mouse models of AD; and to develop a preclinical testing pipeline.

Goals of the MODEL-AD Consortium are:

- Develop the next generation of in vivo AD models based on human data
- Institute a standardized and rigorous process for characterization of animal models
- Align the pathophysiological features of AD models with corresponding stages of clinical disease using translatable biomarkers
- Establish guidelines for rigorous preclinical testing in animal models
- Ensure rapid availability of animal models, protocols and validation data to all researchers for preclinical drug development

Contact us:

- IU/JAX/PITT
- UCI
Assessing Human-relevant Omics Signatures in Mouse Models

AMP-AD Human LOAD Transcriptomic Modules
MODEL-AD: Enabling the NIA Translational Research Pipeline for AD/ADRD

Clinical Data

LOAD Models

Translational Phenotypes

Drug Testing

Clinical Data
- ‘Omics
- Neuropathology
- Neuroimaging
- Fluid Biomarkers

LOAD Models
- Neuroimaging
- Neuropathology
- Fluid Biomarkers

Translational Phenotypes
- ‘Omics
- Neuropathology
- Neuroimaging
- Fluid Biomarkers

Drug Testing
- Territo Lab
- Rizzo Lab
- STOP-AD Portal
- Neuromarkers
- In vivo target engagement
- In vivo PK
- In vivo Efficacy
- Processor
- Drug
MODEL-AD: Models Distribution across Academia and Industry

https://www.model-ad.org/data-and-resources/
Over 70 mouse strains/models available
Prioritized Targets

Admin and Data Core
Assay Development and High-Throughput Screening Core
Structural Biology Core
Chemical Biology and Medicinal Chemistry Core

Target Enabling Packages

Assay Development

In vitro/In vivo Chemical Probes

Open distribution of knowledge, data and target enabling tools

https://treatad.org
**Components of Target Enabling Packages**

**Informatics Packages**

- AD Risk Scores
- AD Knowledge Portal

**Reagent Packages**

- Knockout cell line
- Validated antibodies
- Purified Protein
- Biochemical & Biophysical assays
- Crystal structures
- Target engagement assays
- Functional modulators (biologic or chemical)

Available reagents for enabled targets

[adknowledgeportal.org](http://adknowledgeportal.org)

Available for all genes/proteins

[agora.adknowledgeportal.org](http://agora.adknowledgeportal.org)

---

Axtman AD, et al., Open drug discovery in Alzheimer's disease, Alzheimer’s & Dementia TRCI, 2023
### Current TEP Portfolio

<table>
<thead>
<tr>
<th>Gene</th>
<th>Bioinformatic analysis</th>
<th>Antibody validation</th>
<th>Protein constructs</th>
<th>Crystal structure</th>
<th>Biochemical assay</th>
<th>Biophysical assay</th>
<th>Cell-based assay</th>
<th>Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SERP1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CAPN2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MBN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CD44</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MDK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SDC4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ARHGEF2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IFH1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DHX58</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DOX1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TICAM1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FCE1G</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NDUFS2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CTSH</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CNN3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GPNMB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLE1C</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROX1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROX6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RABEP1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>STK4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C4A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SMC1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SNX32</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EPHX2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FCE1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ECE1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>APDE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PTN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DAG1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OPRT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HTRA1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**KEY**

- Available on the Portal
- Complete
- In progress
- Not applicable/planned

[Target Portfolio Dashboard](https://treatad.org/)
NIA ACTDRx AD Program: Leveraging AMP AD Deliverables to Accelerate Drug Repurposing and Combination Therapy Development

Established in 2018

- Robust and highly cross-disciplinary program bringing together teams with deep expertise in data science, AI/ML, multiscale modeling working on ADRD and other chronic diseases

- Leveraging AMP AD data in combination with other data resources (ADNI, ADSP, LINCS, EHR, Medicare/Insurance claims) to enable data-driven drug repurposing and combination therapy development for AD

- Identified dozens of repurposable drug candidates through data-driven predictions and resulted in one Phase 1/Phase 2 clinical trial for AD

Work in progress: Incorporating drug predictions and supporting evidence in Agora to enable data driven prioritization of repurposable drug candidates.
Open Science Programs Enabling a Precision Medicine Approach to Drug Development for AD

AMP-AD
M²OVE-AD
CLEAR-AD
AGMP
www.ampadportal.org

TREAT-AD
MODEL-AD
Psych-AD
Resilience-AD
ACTDRx AD
MARMO-AD

AD Knowledge Portal
AD Knowledge Portal
www.adknowledgeportal.org

Agora
agora.adknowledgeportal.org

Model AD Explorer
www.modeladexplorer.org
AD Knowledge Portal
Evolving to Enhance Interoperability and Better Support a Diverse Set of End-users

- AMP-AD Portal launch (Synapse)
- AD Knowledge Portal relaunch (JavaScript)
- Download cart
- News site
- New Study detail pages
- New homepage
- Technical vignettes
- Publication datasets
- Redesigned Project pages

2015

- Faceted file browser
- Keyword search
- Access restriction icons
- Data access request in Portal
- Analytical workspace

2019

2020

- New Improvements
- Cohort builder
- Improved high-level navigation
- Full site search
- Interoperability with other ADRD data repositories and AMP platforms

2021

- Documentation/Help site
- Experimental Tools Redesign
- Animal Models Redesign

2023
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)
AMP AD

Noel Burtt (Broad)
AMP CMD

Sweta Ladwa (NHLBI)
AMP HF

Matt Bookman (Verily)
AMP PD

Anna Greenwood (Sage)
AMP data landscape
AMP SBI goals and challenges
The AMP® Common Metabolic Diseases Knowledge Portal: Approach & Impact

February 6, 2024
Noël Burtt
cmdkp.org
ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES GENETICS beta

TCF7L2

Uniprot Summary: Participates in the Wnt signaling pathway and modulates MYC expression by binding to its promoter in a sequence-specific manner. Acts as a repressor in the absence of CTNNB1, and as an activator in its presence. Activates transcription from promoters with several copies of the Tcf-binding motif 5'-CCTTTGATC-3'. In the presence of CTNNB1, TLE1, TLE2, TLE3 and TLE4 repress transactivation mediated by TCF7L2/TCF4 and CTNNB1. Expression of dominant-negative mutants results in cell-cycle arrest in G1, necessary for the maintenance of the epithelial stem-cell compartment of the small intestine.

Variants and associations

Explore variants within 100kb of TCF7L2

Click on a number below to generate a table of variants associated with type 2 diabetes in the following categories:

<table>
<thead>
<tr>
<th>data type</th>
<th>sample size</th>
<th>total variants</th>
<th>genome-wide significant variants</th>
<th>locus-wide significant variants</th>
<th>nominal significant variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS</td>
<td>69,033</td>
<td>230</td>
<td>39</td>
<td>58</td>
<td>123</td>
</tr>
<tr>
<td>eQTL chip</td>
<td>79,854</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>eQTL sequence</td>
<td>16,760</td>
<td>161</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Variants within 500kb of TCF7L2 are also genome-wide significantly associated with:
- two-hour glucose
- fasting glucose
- fasting proinsulin

Explore significant variants with IGV

9 Datasets, 25 traits
Type 2 Diabetes Knowledge Portal

Type 2 Diabetes Knowledge Portal

Providing data and tools to promote understanding and treatment of type 2 diabetes and its complications

Search gene, variant, region or phenotype

examples: SLC30A8, rs13266634, chr9:21,940,000-22,190,000, Type 2 diabetes

Make genetic association data available to the wider research community
Cell Metabolism

The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits

Graphical abstract

Authors
Maria C. Costanzo,
Marcin von Grothuss,
Jeffrey Massung, .... Michael Boehnke,
Noël P. Burtt, Jason Flannick

Correspondence
burtt@broadinstitute.org (N.P.B.),
jason.flannick@childrens.harvard.edu (J.F.)

In brief
The Type 2 Diabetes Knowledge Portal (T2DKP) is an innovative resource that democratizes access to human genetic and genomic data. In this issue, Costanzo et al. describe how both novice and expert users can use the T2DKP in their research.
50 countries represented

350 contribution directly in collaboration!
The resource rests on a biological model.
Integrate data & present the knowledge

- Association statistics
- Reference chromatin state
- Chromatin capture
- Networks
- Cellular models
- Meta-analysis
- Transcription factor binding sites
- Gene expression
- Pathways
- Fine mapping
- eQTLs
- Gene function
- Animal models
- Variant effect predictors
- Regulatory element prediction
- Gene prioritization
- Gene set enrichment
- Tissue-specific enrichment
- eQTL colocalization
- Chromosome contact prediction

Summary Representations
Scientific Questions
Results & Knowledge
CMDGA aggregates functional genomics data from tissues/cell types relevant to common metabolic and other complex diseases to annotate (i) genes and gene regulatory elements and (ii) genetic effects on gene regulatory activity.

**Gene regulatory elements**
- Candidate CREs
- Target gene pred.
- Histone mod.
- Chromatin states
- Accessible chrom.
- TF binding sites
- Gene expression
- Differential activity
- eQTLs
- Variant effects
- Sc embeddings
- Other QTLs
- Perturbations

**Gene activity and function**
- Target gene links
- 3D interactions
- Co-accessibility
- Activity-by-contact
- Element screens
- Gene expression
- Gene function screens
- Gene perturbations

**Data and meta-data for donors, experiments, embeddings, models, annotations, etc.**
- Analytical pipelines for standardized processing of data
- Web portal and APIs for data query and access
- Visualization tools to interact with single cell profiles

**Annotations in CMDGA:**
- Candidate CREs
- Target gene pred.
- Histone mod.
- Chromatin states
- Accessible chrom.
- TF binding sites
- Gene expression
- Differential activity
- eQTLs
- Variant effects
- Sc embeddings
- Other QTLs
- Perturbations
Integrate genetic & genomics data to generate hypotheses about molecular or cellular function

- Meta-analysis
- Variant to Gene Predictions
- Target Gene Links
- Tissue Enrichment
- Pathway Enrichment
- Genetic Correlations

Analysis Platform
Impact: Flexible Software Platform
Impact: On-demand research

Without CMDKP Knowledgebase

1. Identify studies at GWAS Catalog
2. Download summary statistics
3. Metaanalyze
4. Run annotation pipeline
5. Identify predicted impactful SNPs in gene of interest
6. Run statistical analyses
7. Weight of evidence for role of gene in disease

Gene expression results
- Gene 1
- Gene 2
- Gene 3
- Gene 4

Model organism results

List of target genes to investigate

Literature
Impact: Novel approaches accessible

Gene-level associations for FTO

Click the tabs below to view gene-level association scores for FTO based on the HuGE Calculator, MAGMA analysis of common variants, or burden tests of rare variants.

<table>
<thead>
<tr>
<th>HuGE Scores</th>
<th>Common variant associations</th>
<th>Rare variant associations</th>
</tr>
</thead>
</table>

HuGE Scores

HuGE (Human Genetic Evidence. Donbros et al. 2022) scores quantify genetic support for involvement of FTO in the diseases and traits available in the Portal, based on several kinds of human genetic results. See the HuGE Calculator documentation for more details.

Customized analyses may be performed on the HuGE Calculator page.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Group</th>
<th>Common Variation Bayes Factor</th>
<th>Rare Variation Bayes Factor</th>
<th>HuGE Score</th>
<th>Evidence Range</th>
<th>HuGE Calculator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tnglcyctes</td>
<td>LIPIDS</td>
<td>45.00</td>
<td>1.50</td>
<td>67.52</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>Weight</td>
<td>ANTHROPOMETRIC</td>
<td>45.00</td>
<td>1.01</td>
<td>45.30</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>ANTHROPOMETRIC</td>
<td>45.00</td>
<td>1.00</td>
<td>45.02</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>Obesity</td>
<td>CARDIOVASCULAR</td>
<td>45.00</td>
<td>1.00</td>
<td>45.00</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>Childhood obesity</td>
<td>ANTHROPOMETRIC</td>
<td>45.00</td>
<td>1.00</td>
<td>45.00</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>RENAL</td>
<td>45.00</td>
<td>1.00</td>
<td>45.00</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>White blood cell count</td>
<td>HEMATOLOGICAL</td>
<td>45.00</td>
<td>1.00</td>
<td>45.00</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
<tr>
<td>Glucose</td>
<td>CARDIOVASCULAR</td>
<td>45.00</td>
<td>1.00</td>
<td>45.00</td>
<td>Very Strong</td>
<td>Open</td>
</tr>
</tbody>
</table>
Impact: Expert Curated Content

150 curated predicted effector gene lists with evidence source
The resource over time

- LocusZoom
- GAIT
- Datasets page
- First genomic annotations
- Federation with EBI
- Region page
- Variant PheWAS
- Variant forest plot
- Bottom line integrative analysis
- First effector gene list
- GREGOR tissue enrichment
- User interface major update
- Genomic Region Miner
- Gene Finder
- MAGMA gene-level analysis
- KP Labs page
- LD clumping
- Lunaris
- HuGE Calculator
- Signal Sifter
- Noncoding GAIT
- Variant Sifter
- MAGMA pathway analysis
- Genetic correlations
- Transcript-specific rare variant associations
- Transcript-level burden tests
- Ancestry-specific genetic associations
- HuGE Calculator encoded
- Gene expression viewer
- LDSR implemented
- Gene Sifter
- Top line integrative analysis
- First functional gene list
- GREGOR tissue enrichment
- User interface major update
- Genomic Region Miner
- Gene Finder
- MAGMA gene-level analysis
- KP Labs page
- LD clumping
- Lunaris
- HuGE Calculator
- Signal Sifter
- Noncoding GAIT
- Variant Sifter
- MAGMA pathway analysis
- Genetic correlations
- Transcript-specific rare variant associations
- Transcript-level burden tests
- Ancestry-specific genetic associations
- HuGE Calculator encoded
- Gene expression viewer
- LDSR implemented
- Gene Sifter

The resource over time
Expand the platform to an NHGRI Community Genomic Resource

[Image of Community Knowledge Portals]

[Image of All of Us Function Knowledge Portal]

a2fkp.org
Our philosophy & approach

• **Collaborate** directly with disease research communities to understand data, analytical results required to represent *knowledge*

• **Onboard** precomputed results & generating summary representations from contributors

• **Incorporate** external resources

• **Encode** analytical methods

• **Integrate** results, *accessible* in an intuitive, browsable, open-access, web portal
Multi-Site Team

Broad/DCC
- MacKenzie Brandes
- Maria Costanzo
- Marc Duby
- Drew Hite
- Quy Hoang
- DK Jang
- Ryan Koesterer
- Fahrisa Islam Maisha
- Annie Moriondo
- Trang Nguyen
- Phebe Olorunfemi
- Oliver Ruebenacker
- Alex Shilin
- Patrick Smadbeck
- Noël Burtt
- Jason Flannick

UM
- Ryan Welch
- Laura Scott
- Michael Boehnke

EBI
- Laura Harris
- Yue Ji
- Aoife McMahon
- Thomas Keane

UCSD/CMDGA
- Parul Kudtarkar
- Ying Sun
- Kyle Gaulton

Logos:
- All of Us
- HuGeAMP
- KPN
- National Institute of Diabetes and Digestive and Kidney Diseases
- FNIH
- Common Metabolic Diseases Genome Atlas
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)
AMP AD

Noel Burtt (Broad)
AMP CMD

Sweta Ladwa (NHLBI)
AMP HF

Matt Bookman (Verily)
AMP PD

Anna Greenwood (Sage)
AMP data landscape
AMP SBI goals and challenges
NHLBI’s BioData Catalyst® for the AMP-HF

Sweta Ladwa
- Chief, Scientific Solutions Delivery Branch
- Scientific Program Director, BDC Data Management Core
Information Technology & Applications Center
NHLBI
Topics

- AMP-HF and NHLBI’s BioData Catalyst® Overview
- Milestones Achieved
- Path forward
AMP-HF

AMP HF is the 10th AMP project to be launched (2022) – it’s the new kid on the block!

- Leveraging NHLBI’s HeartShare program which aims to conduct large-scale analysis of phenotypic data, images, and omics from patients with heart failure with preserved ejection fraction (HFP EF) and controls in order to characterize mechanisms of disease and identify therapeutic targets.
  - Create a rich repository of data, omics, and images from multiple large scale cardiovascular disease (CVD) epidemiology studies, multicenter heart failure (HF) clinical trials, and other studies to facilitate open data science for the investigation of heart failure with preserved ejection fraction (HFP EF).
  - Define subtypes of HFP EF, and develop algorithms (e.g., biomarkers, image analyses) for identifying the subtypes.
  - Discovery and prioritization of targets in mechanistic pathways for diagnosis, risk assessment, and development of new therapies.
  - Establish a large, electronic health record (EHR) based cohort of HF patients and comorbidity-matched controls who can be enrolled in a registry for future studies.
  - Establish a cohort of deeply-phenotyped patients who can support target validation and facilitate proof-of-concept treatment trials for HFP EF.
  - Train new HF data scientists proficient in multi-omics, machine learning/augmented intelligence techniques.
AMP-HF and NHLBI’s BioData Catalyst

- Much of the retrospective omics and phenotypic data HeartShare and the AMP plan to use are already in NHLBI’s BioData Catalyst ecosystem
  - Since BioData Catalyst (BDC) hosts a data repository as well as computational tools and platforms for advanced data science, a new portal for AMP-HF was not necessary to build.
NHLBI BioData Catalyst

The **mission** is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

The **vision** is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.
NHLBI’s BioData Catalyst® Overview
NHLBI’s BioData Catalyst® Overview

BioData Catalyst provides an infrastructure that allows researchers to bring their own tools and workflow to the data, as well as use published tools and workflows within the ecosystem with ease.

- Collaborative, Multi-modal Analytic Capabilities in BDC Workspaces (Seven Bridges and Terra) to support AMP investigators at all stages of their research lifecycle:
  - Analytic workspaces (Jupyter, R, Python, SAS)
  - GWAS (using GENESIS or HAIL)
  - Variant Calling using GATK or TOPMed (WES/WGS Data)
  - Image visualization and analysis including Machine Learning models
  - App Development
  - App Hosting
  - Pipeline Development (CWL or WDL, Nexus coming soon)
  - Import and publish apps and workflows to/from dockerized apps (e.g. Dockstore)
- Search:
  - Search UI – Genotype and Phenotype
  - Search API
  - Synthetic Cohort Generation
  - Semantic meta-data search
BioData Catalyst® and Interoperability

NIH Cloud Platform Interoperability (NCPI)
Federate data analysis (bring compute to data with NIH Partner Systems)

Advanced Research Projects Agency for Health (ARPA-H)
Biomedical Data Fabric Toolbox
Support a national, unified consistent layer of data services to work across many different systems and environments

Global Alliance for Genomics and Health
Work to interoperate with international data sources, including UK Biobank and the NIH All of Us Research Program
Milestones Achieved for AMP-HF

- Established an image de-identification protocol for imaging critical to AMP-HF scientific aims
  - Deidentification protocol is for all imaging modalities (e.g. echo, ECG, MRI, etc.)
  - Supporting reading center deidentification and data transfer to NHLBI cloud buckets for ingest to BDC and use by the AMP-HF teams
- Actively exploring options for image data analysis within BDC to support all prongs of the AMP-HF scientific aims
- Data access for most of the longitudinal cohort studies for the teams have been approved, with data starting to be accessible via the BDC Workspaces in Seven Bridges
- For controlled access data from NHLBI cohort studies, as well as other clinical trials, we established a Non-Genomic DAC via dbGaP to maintain access controls in BDC and promote FAIR data
BioData Catalyst® is Continuously Developing

While supporting HeartShare and AMP-HF has been a driver project for BDC, BDC is incorporating new features and capabilities to not only support the AMP-HF teams, but also all HLBS researchers (bolded came directly from AMP-HF)

- Supporting collaboration along the entire research lifecycle continuum
- Adding infrastructure to support imaging AI/ML analysis
- Meeting NHLBI Data Science Program and FAIR data needs (e.g. HeartShare, TOPMed, CureSCi, etc.)
- Meet the needs of researcher compliance with the NIH Final Policy for Data Management and Sharing
- Implementing DataCite DOIs through the NIH ODSS STRIDES program for hosted data
- Investigating data management and analytic applications of Large Language Models (LLMs) for Heart, Lung, Blood, and Sleep researchers
- Expanding support for various data types (metadata, etc.)
  - Single Cell Omics
  - Cell characterization
  - Proteomics
Acknowledgements

- NHLBI Director: Dr. Gary Gibbons
- BDC Executive Sponsor: Dr. David Goff
- BDC Scientific Program Director: Dr. Regina Bures
- BDC DMC Scientific Program Director: Mrs. Sweta Ladwa
- BDC Steering Committee: Dr. Ingrid Borecki, PIs, Consortia
- BDC EEP
- BDC Coordinating Center: Dr. Ashok Krishnamurthy/Dr. Stan Ahalt
- RTI International: Mrs. Becky Boyles
- University of Chicago: Dr. Bob Grossman
- Velsera: Dr. Jack DiGiovanna
- Harvard Medical School: Dr. Paul Avillach
- Broad Institute: Dr. Alisa Manning
- AMP-HF: Dr. Vandana Sachdev
- HeartShare: Dr. Sanjiv Shah

The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst® (BDC) ecosystem.
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)
AMP AD

Noel Burtt (Broad)
AMP CMD

Sweta Ladwa (NHLBI)
AMP HF

Matt Bookman (Verily)
AMP PD

Anna Greenwood (Sage)
AMP data landscape
AMP SBI goals and challenges
AMP PD - Platform Vision
Bring researchers and tools to the data

A new approach

Traditionally, data sharing has involved

- Bringing data to researchers
- Tools

Disadvantages:

- Discourages shared research
- Facilitates collaboration

A Data Biosphere for Biomedical Research

Benedict Paten

Josh Denny (Vanderbilt), David Glazer (Verily Life Sciences), Robert L. Grossman (University of Chicago), Benedict Paten (University of California at Santa Cruz), Anthony Philippakis (Broad Institute)

5 min read  Oct 16, 2017
Bring researchers and tools to the data

A new approach to biomedical research

Traditional approach

Bring data to researchers and build

Challenges
- Data sharing = data copying + poor security
- Bespoke & unsupported tools
- Huge infrastructure needed
- Discourages shared research

Terra’s approach

Bring research to data and share

Integrated Platform

Advantages
- More accessible and secure data
- Shared tool ecosystem
- Decreased cost of storage & compute
- Facilitates collaboration
Bring researchers and tools to the data

- There is overhead in researchers learning a new system, but gains for the community will accrue
  - More accessible and secure data
  - Shared tool ecosystem
  - Decreased cost of storage & compute
  - Facilitates collaboration

Using a public cloud brings global compute power to every researcher, not just those at large institutions.
Build on platform(s) that will keep getting better

- Google Cloud
  - Global team designing and building out new capabilities
  - Vibrant community of companies and open source teams building new software that runs on public cloud platforms
- Terra
  - Team designing and building out new capabilities
  - Direct collaborations with large programs, such as AnVIL bring features like
    - RStudio
    - Galaxy
  - Community contributions of Featured Workspaces with new tools
Operations Workbench == Researcher Workbench

- We run workflows to turn raw data into analysis-ready data
- We run notebooks for quality control and visualization
A Workbench for Direct Data Access

- I've been granted access to 100 TB of WGS and RNAseq data
  - Traditional approach
    - I'm going to contact my IT to requisition storage space
    - I'm going to let them know that I'll be downloading data for a few weeks (or file a ticket requesting that they do it)
    - I'm going to request compute capacity for processing
  - New approach
    - I've got getting started notebooks in my preferred language(s)
    - I've got data loaded in a dataframe in ~5 minutes
    - I can scale up my computation as needed
A Workbench for Direct Data Access

ABOUT THE WORKSPACE

AMP PD - Getting Started Tier-2 - Clinical and Omics Access - Version 4

The purpose of this workspace is to provide getting started information and notebooks for researchers granted access to AMP PD data.
A Workbench for Collaborations

- Researcher at Institution A runs an analysis and wants to share with researcher at Institution B
  - Traditional approach
    - I'll ask our IT team if we can get you a guest account...
    - You can't connect into our network, so...
    - I'll send you my code to run on your systems...
  - New approach
    - Create a workbench account and I'll share my workspace with you
    - You can run on the exact same configuration as me

And extend both narratives to collaborating with a researcher from Institution C...
A Workbench for Collaborations

- We have this same experience time and again
- Example: RNASeq Sex Checks
A Workbench for Collaborations

- RNASEq Sex Checks...

4 different professions, 3 different institutions, collaborating over a few days to validate that there was no large-scale mis-labeling of samples...
What's Next?
Looking forward - AMP PDRD

- Cross analysis with AMP AD identified as a priority
  - Workstreams
    - Data harmonization
      - Identify priority data for harmonization
      - Identify path to harmonize those elements
    - Interoperability
      - Federate access across platforms?
      - Federate access to AMP ADAPT?
      - Centralize a subset of the data for targeted analysis?

Clear synergy with AMP SBI
Thank-you
Data, Data, Everywhere - AMP Portals and Interoperability

Suzana Petanceska (NIA)
AMP AD

Noel Burtt (Broad)
AMP CMD

Sweta Ladwa (NHLBI)
AMP HF

Matt Bookman (Verily)
AMP PD

Anna Greenwood (Sage)
AMP data landscape
AMP SBI goals and challenges
AMP Data Landscape &
AMP SBI Goals and Challenges

2/6/2024

Anna Greenwood, PhD
Sage Bionetworks
• Many diseases studied in AMP programs are characterized by inflammatory processes

• Goal of AMP-SBI is to identify shared and distinct mechanisms active in subsets of patients across multiple diseases

• This project seeks to leverage the breadth of existing datasets by integrating data across AMP projects
How do we enable integration of data?

Data

Infrastructure

Governance
### Data types and source tissues differ across AMP projects

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AMP Data Types</strong></td>
<td>Raw and processed genomic, <strong>transcriptomic</strong>, epigenomic, proteomic, metabolomic, clinical and demographic data</td>
<td>Raw and processed genomic, transcriptomic, epigenomic, proteomics, mass cytometry, histology, clinical and demographic data</td>
<td>Summarization level genetics data. Raw data through dbGaP/EGA Raw &amp; processed genomic, transcriptomic, epigenomic data available</td>
<td>Raw and processed genomics, transcriptomics, proteomics, clinical and demographic data</td>
<td>Raw and processed genomic, EEG, MRI, Cognitive data, Personal tracking device data (actigraphy, phones), clinical and demographic data</td>
<td>Raw and processed genomic, transcriptomic, epigenomic, proteomics, metabolomics, clinical, demographic data, sleep (actigraphy, PSG, etc.), imaging (echo, ECG, MRI, etc.), wearables, EHR</td>
</tr>
<tr>
<td><strong>Data Sources</strong></td>
<td>Human, animal models, in-vitro model systems</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
</tr>
<tr>
<td><strong>Targeted Tissue</strong></td>
<td>Brain, blood, CSF</td>
<td>Synovium, kidney, urine, blood, skin</td>
<td>Pancreas, adipose, heart, liver, kidney, skeletal muscle, epithelium, blood</td>
<td>Brain, Blood, CSF</td>
<td>Blood (genotypes)</td>
<td>Cardiac, skeletal muscle, adipose tissue</td>
</tr>
</tbody>
</table>
Data repositories and infrastructure differ across AMP projects

AMP-RA/SLE
AMP-SCZ
AMP-PD
AMP-AD
AMP-T2D
AMP-CMD
AMP-HF
## Governance and data access differ across AMP Projects

<table>
<thead>
<tr>
<th>Access Tiers</th>
<th>AMP AD</th>
<th>AMP PD</th>
<th>AMP RA/SLE/AIM</th>
<th>AMP T2D/CMD</th>
<th>AMP SCZ</th>
<th>AMP HF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open</strong></td>
<td>Study descriptions, methods, summary stats, file level metadata</td>
<td>Study descriptions, methods, summary stats</td>
<td>Study descriptions, methods, summary stats, file level metadata</td>
<td>All content is open access</td>
<td>Study descriptions, data dictionaries</td>
<td>Study and phenotype descriptions</td>
</tr>
<tr>
<td><strong>Registered</strong></td>
<td>Animal models, human summary results</td>
<td>De-identified clinical data and human summary results</td>
<td>Human summary results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Controlled</strong></td>
<td>Individual level human data</td>
<td>Individual level human data</td>
<td>Individual level human data</td>
<td>Individual level human data</td>
<td>Individual level human data</td>
<td>Individual level human data</td>
</tr>
</tbody>
</table>
Varied end users and use cases across AMP projects

Computational biologist

Experimental biologist

And: patients, patient advocates, clinicians, drug developers, etc....
Potential for reusability of data varies by data and user types

### Gene expression
- **Raw data**: fastq files
- **Processed data**: transcript counts
- **Analytical results**: differential gene expression in brain in AD case vs controls

### Genetic variants
- **Raw data**: fastq files
- **Processed data**: variant calls
- **Analytical results**: GWAS association statistics
AMP programs have developed solutions for varied data and user types

**AD Knowledge Portal**
Raw and processed data

**Agora**
Curated results explorer
AMP programs have developed solutions for varied data and user types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Raw data</th>
<th>Processed data</th>
<th>Analytical results</th>
<th>Curated results explorer</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td></td>
<td>AD Knowledge Portal</td>
<td>Agora</td>
<td></td>
</tr>
<tr>
<td>RA/SLE/AIM</td>
<td></td>
<td>ARK The Autoimmune and Related Disease Knowledge Portal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D/CMD</td>
<td></td>
<td>dbGaP GENETICS ARCHIVES, EUROPEAN GENOME-PHENOME ARCHIVE, CMDKP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td></td>
<td>AMP PD Portal</td>
<td>AMP PD Target Explorer</td>
<td></td>
</tr>
<tr>
<td>HF</td>
<td></td>
<td>HeartShare DCC</td>
<td>BioData CATALYST</td>
<td></td>
</tr>
</tbody>
</table>
Taking steps toward an integrated AMP data ecosystem

Interoperability from a user perspective:

1. Easily find compatible data and results from different systems
2. Seamlessly access data and results from different systems
3. Analyze all data in a common environment
Use Cases from AMP-SBI Pilot Projects

- 6 pilot projects funded

- Projects interrogating data across multiple AMP projects
  - Requirement: inclusion of non-neurodegenerative disease

- Range of projects including:
  - Evaluation of shared inflammatory signatures across diseases
  - Evaluation of shared genetic association across diseases and immune traits
  - Integration of single cell/nucleus transcriptomic data from tissue-derived immune cells (central and peripheral)
Use Cases from AMP-SBI Pilot Projects

AD: Parahippocampal Gyrus M153

- Parahippocampal Gyrus M153
- Up-regulated in AD

T2D: Adipose Tissue M6

- Adipose Tissue M6
- Down-regulated in T2D
- Up-regulated in T2D

Slide from Bin Zhang, Mt. Sinai
Focus of AMP ADAPT: Create a federated, interoperable AMP data ecosystem and probe multi-omics data across multiple diseases.

Cross-AMP working groups actively developing proposed project aims and scope:

Aim 1: Data and platform interoperability

Aim 2: Data governance

Aim 3: Enabling analysis of shared and distinct pathways of inflammation
Thank You!

Interoperable AMP Data Ecosystem Survey

• Your input is appreciated to help AMP SBI understand how AMP data can benefit you.

• Please complete this short 3-minute survey using the QR code or at https://rb.gy/m16m6z