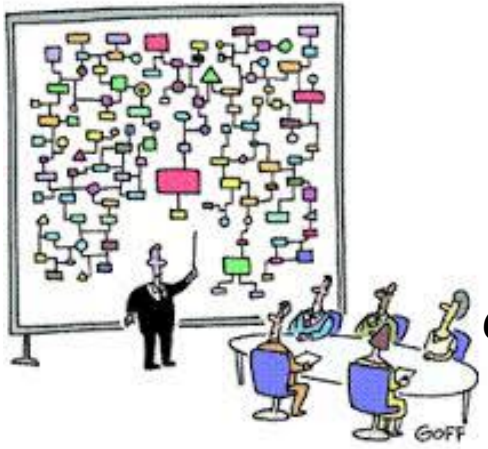
BIOMEDICAL BIG DATA IN CLEVELAND



"And that's why we need a computer."

Jonathan L. Haines, Ph.D.

Chair, Epidemiology & Biostatistics

Director, Institute for Computational Biology

Case Western Reserve University

September 22, 2015

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Department Epidemiology & Biostatistics

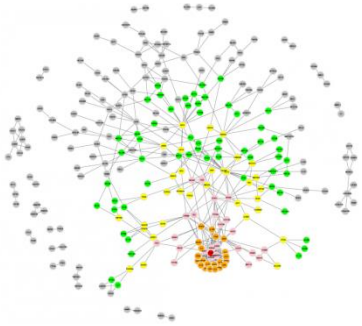
Outline

- **Introduction**

- **Capturing the data**

- **Analyzing the data**

- **Connecting the data**



Outline

- **Introduction**

- Capturing the data

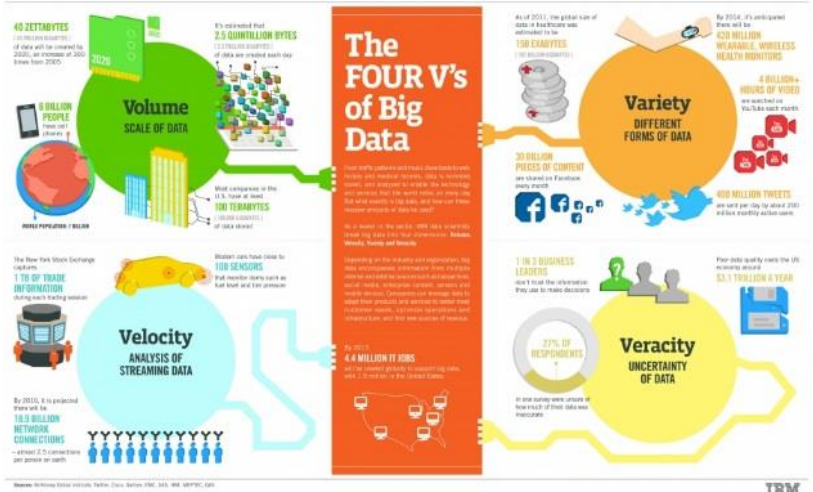
- Analyzing the data

- Connecting the data

What is Big Data?

- The Description
- The Four V's
  - Volume (how much)
  - Velocity (how fast)
  - Variety (how different)
  - Veracity (how good)

- The Challenge
- The Four C's
  - Creation
  - Collection
  - Connection
  - Context



What is Big Data?

My Bottom Line: Any data that doesn't easily fit into Excel!

Big Data Examples

- Data Sources
  - Google searches
  - Grocery store purchases
  - Facebook likes
  - Tweets
  - Cell phone usage
  - Car tracking by Insurance companies
  - Governmental records
  - Anything the NSA is interested in
- Data Analysis
  - Algorithms are often deceptively simple
    - Correlation
    - Regression
    - Decision trees
  - But not always
    - Support Vector Machines
    - Neural Networks
    - Natural language processing

Biomedical Big Data in Research
The Potential

- Improve the description of the natural history of disease
  - Variability and clustering of symptoms
  - Penetrance of genetic mutations
- Discover new disease associations with biomarkers
- Better understand the underlying physiology of a trait
- Better understand the influence of social and physical environment
- Identify targets for drugs
- Recover shelved drugs
- Develop new and refine existing therapies
- Identify the weak points in health care delivery
- Refine the processes for healthcare delivery
- Use this information to improve healthcare delivery to individuals
  - Precision medicine
- Use this information to improve the health of communities
  - Move the needle

Big Data in Biomedicine

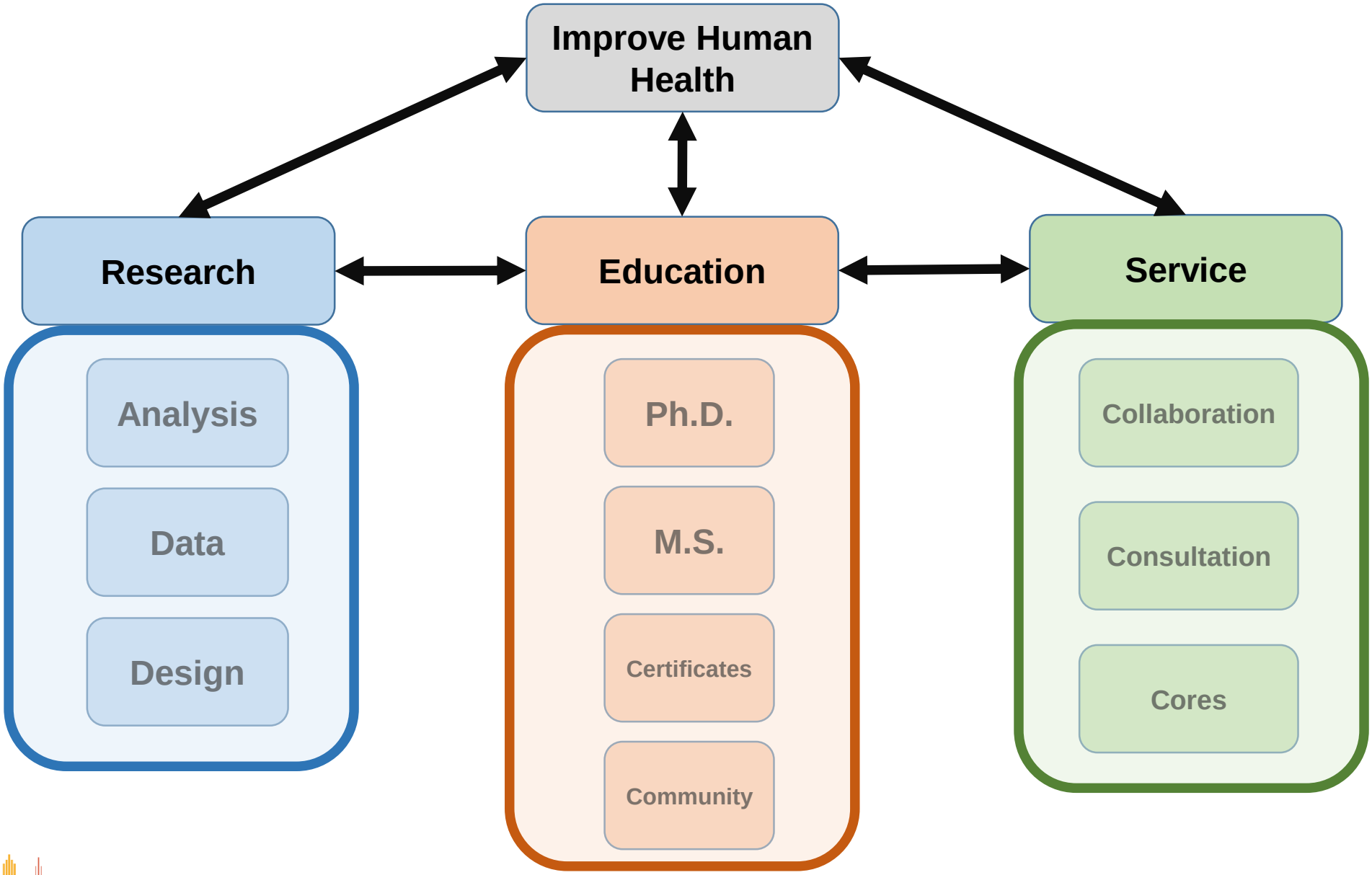
Types

- Electronic Health Records
- Biological data
  - Genomics (DNA sequence)
  - Clinical laboratory measures
  - Proteomics data
  - Metabolomics data
  - Microbiome data
  - Imaging data
- Administrative (billing) data
- Governmental data
- Physical environmental exposure data
- Social environment data

Challenges

- How do we capture and store these data?
- How do we standardize these data?
- How do we normalize and interconnect these data?
- How do we integrate these data?
- How do we mine these data?
- How do we interpret these data into knowledge?
- How do we apply this knowledge to improve health?

Epidemiology & Biostatistics



Outline

- Introduction

- **Capturing the data**

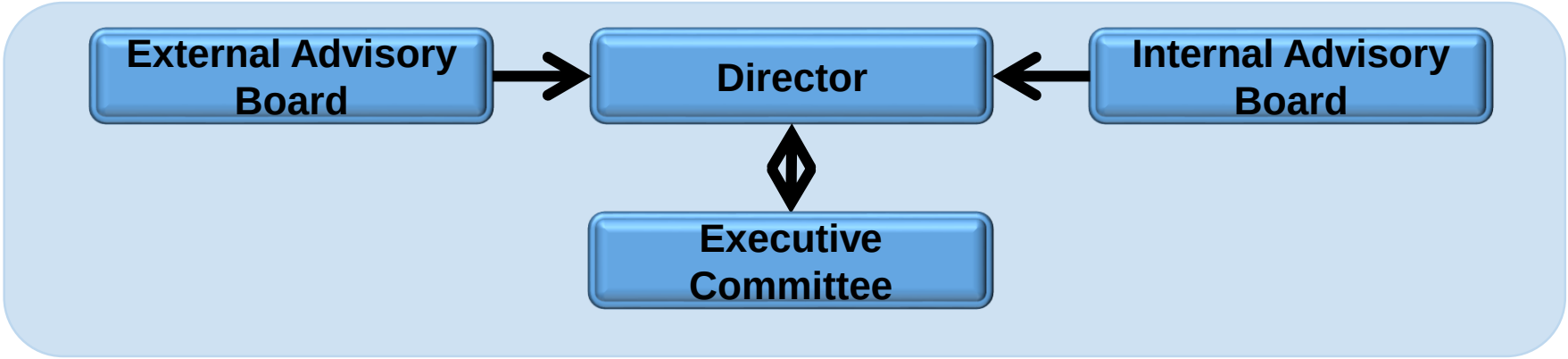
- Analyzing the data

- Connecting the data

Institute For Computational Biology

- **Mission**
  - Advance our fundamental knowledge of human biology through the use of computational methods on big and diverse datasets
  - Promote the translation of this knowledge into better diagnosis, prognosis, treatment, prevention, and delivery
- **Approach**
  - Leverage Cleveland-wide catchment of clinical, epidemiological, biological, socioeconomic, and administrative data in ethnically and economically diverse populations
  - Integrate disparate data types across datasets
  - Promote collaboration through data sharing
  - Enable and encourage innovative analytical approaches to these datasets
  - Advance and enhance educational opportunities
- **Collaborative Funding**
  - Case Western Reserve School of Medicine
  - Cleveland Clinic Foundation
  - University Hospitals
  - MetroHealth (*MOU in progress*)

Institute for Computational Biology



Research Activities

- Program in Research Applications
- Program in Computational Methods

Educational Activities

- Training Grants
- Training Programs
  Cert., MS, Ph.D
  Data Science
- Symposia
- Roadshows

Core Activities

- ***S**afely **H**eld **E**lectronic **D**ata (SHED)*
- ***CLE**veland **A**rea **R**esearch **P**latform for **A**dvancing **T**ranslational **H**ealthcare (CLEARPATH)*
- Computational Biology Core
- Bioethics Core

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Department Epidemiology & Biostatistics

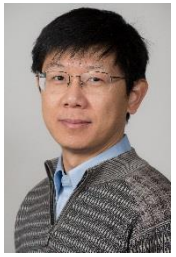
ICB Personnel

Jonathan Haines
Director

Jill Barnholtz-Sloan
Associate Director
Translational Informatics

Will Bush
Assistant Director
Computational Methods

Dana Crawford
Assistant Director
Population & Diversity Research



Chun Li
Assistant Director
Educational Programs

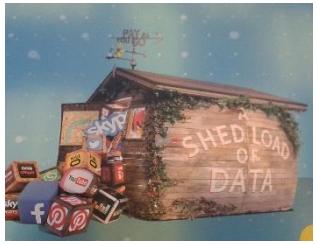
Translational Informatics Team
Patrick Mergler
Mike Warfe
Mark Herron
Bob Lanese
Kellie Bruening
John Turnbull
Devin Tian
Sunah Song
Dan Baechle

Internal Advisory Board
Pam Davis (CWRU)
Fred Rothstein (UH)
Tom Hamilton (CCF)
Mark Chance (CWRU)
Stan Gerson (UH)
Mikkael Sekeres (CCF)
John Foley (UH)
Sue Workman (CWRU)

Core Services
Ricky Chan
Aaron Goldenberg

ICB Infrastructure
Data Capture and Management

- **Investigator Initiated Studies (SHED)**
  - Human subjects (IRB-approved)
  - Legacy, ongoing, and new projects
  - Project-specific focused and detailed data collection
  - Project-specific data structures and queries
  - Identification of commonalities/datasets across individual studies



- **Biomedical Big Data warehouse (CLEARPATH)**
  - Provide a platform to capture and integrate multiple streams of data
  - Provide access and resources to
    - Query the data
    - Develop datasets for research
    - Analyze the data and interpret the results

Safely Held Electronic Data (SHED)
Research Project Management

- **Provides a standardized platform to support PI-initiated research studies**
  - Enterprise level integrated software infrastructure
  - Unique ID for each participant
  - Data housed in a secure data center
  - Data management expertise
- **Advantages**
  - Greatly reduces the risk of data breaches
  - Standardizes and harmonizes data collection
  - Allows real-time exploration of aggregate de-identified data across studies
  - Increases collaborative opportunities
  - Maximizes value of existing/ongoing generated and collected data
- **Governance structure and SOPs to assure appropriate access and management of data**

CLEARPATH
Biomedical Big Data Warehouse

- **Comprehensive database with Limited Data Set contribution from all (3) major Cleveland healthcare systems**
  - Create connected data (Biospecimens, EHR clinical phenotype, 'Omics data)
  - Create a synthetic record for each person ***across systems***
  - Minimizes patient duplication across institutions
- **Robust and flexible data model**
  - Manage a wide range of disease biomarkers, disease classification/ stratification/ staging systems
  - Include processed 'omics, imaging, etc. data
- **Researcher friendly portal**
  - Enables feasibility and preparatory pilot queries
  - Allows research cohort discovery across the aggregate data

Integrated Architecture

Hospitals | **CWRU** | **Investigators**

CCF Network

Biorepositories, EMR/ EDW, Department Registries

IDEAS Server Integrated Data Exchange & Aggregation System

Limited Data

UH Network

Biorepositories, EMR/ EDW, Department Registries

IDEAS Server Integrated Data Exchange & Aggregation System

Limited Data

Metro Network

Biorepositories, EMR/ EDW, Department Registries

IDEAS Server Integrated Data Exchange & Aggregation System

Limited Data

CWRU Secure Research Environment

CLEARPATH

Institutional Repository

Institutional Repository

Institutional Repository

Subject Record Merge

Aggregate Data

Patient Cohort Discovery

Advanced Cohort Analytics

Deep Molecular Analysis

DUA HIPAA Waiver

DUA HIPAA Waiver

DUA HIPAA Waiver

Role Based Access & Auditing

Master Data Management

Limited Data

SHED

IDEAS Server Integrated Data Exchange & Aggregation System

Oncore

LabMatrix

Custom Study Database

Researchers

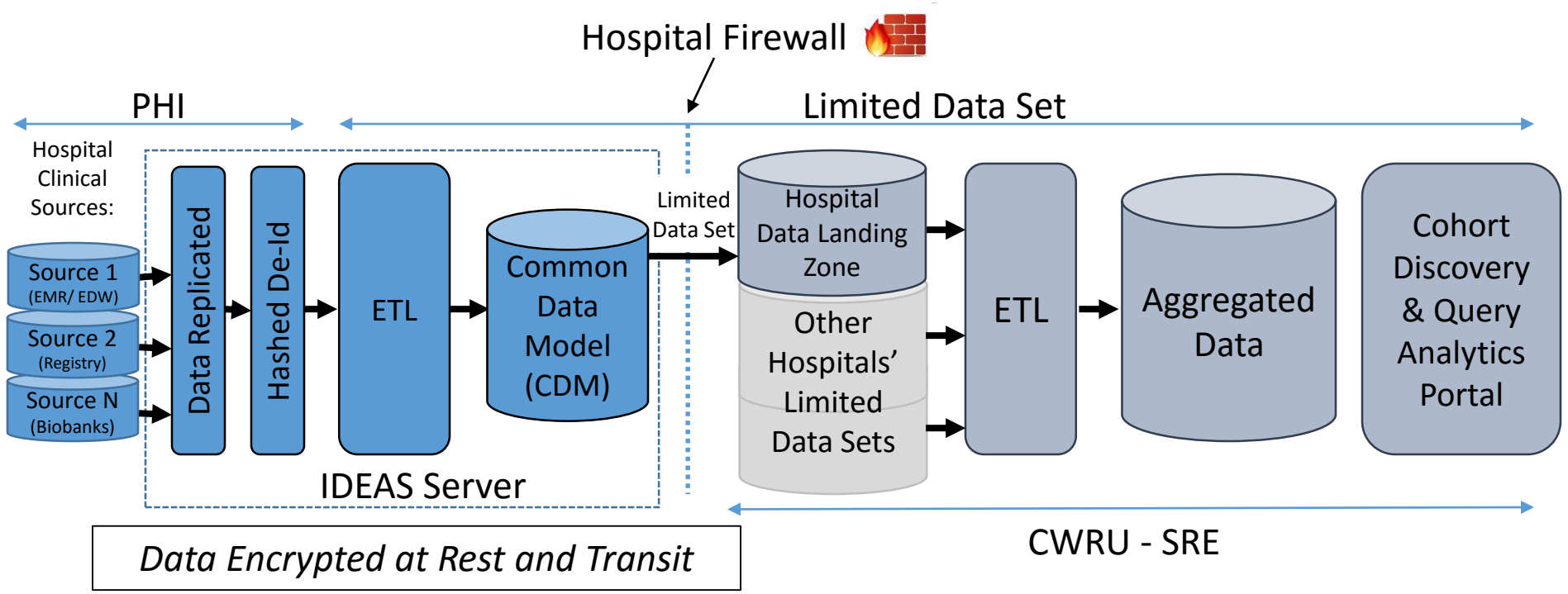
Feasibility and Preparatory Research Query

IRB Approval

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Department Epidemiology & Biostatistics

Data Flow From A Hospital



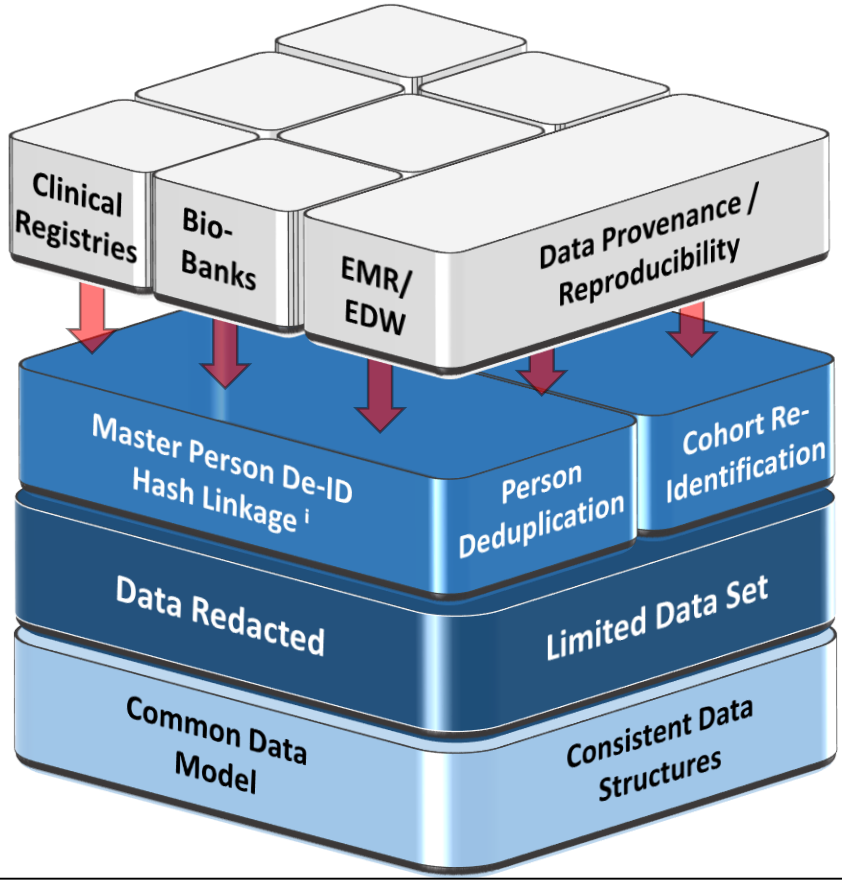
ICB's IDEAS Servers
Reside Behind Hospitals' Firewall

PHI

Limited Data

Source Data Acquisition

Profiling, Cleansing, Indexing, Transparency, and Normalization



Clinical Institution and ICB co-managed

ICB managed, under clinical institution governance

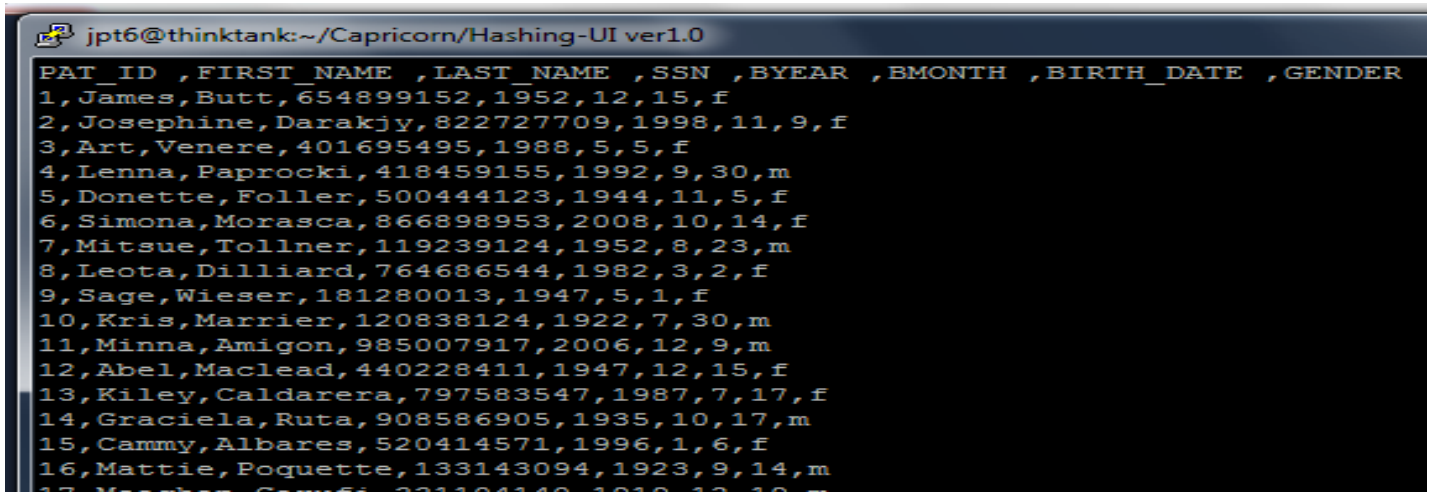
[i] In accordance with HIPAA Privacy Rule 164.514(c) and 164.514 (b) (1).
Expert review letter available upon request

De-identification of Individual Records
HealthLNK: Hashed De-Identification

- **A technology solution that:**
  - **Provides a federated, hashed form of master person indexing across a research network without sharing PHI outside of a hospital firewall**
  - **Enables participating institutions to perform person re-identification upon approved requests**
  - **Identifies duplicate individuals across institutions; enables aggregation of an individual's data across systems: a synthetic medical record**
- **Privacy expert reviewed – meets HIPAA HITECH law**
  - **Only shares de-identified data outside hospital network**
  - **De-identification & re-identification in accordance with HIPAA Privacy Rules 164.514 (b)(1) and 164.514(c)**
- **Already in use supporting PCORI Capricorn research network in Chicago**
  - **11 clinical institutions sharing data (including 2 VA hospitals)**
  - **HealthLNK approved by VA's central office in DC**

How Does This Look in Real-Life?

PHI Input: *(File or database inside Hospital Firewall)*



** Simulated Data*

Hashed De-ID Output: *(Data sent out of hospital)*

HASHED_ID	FNLNDOB	LNFNDOB
CD06F17C38939C5AD5717751461495EA	2d03f230bf2afe70164e549e4836c898644654e1d16b5dc331...	83c8bd5bc7ed
1EAF460C7C29B92CA5EABEEE26DE2CA6	c3e768cf3dde4a27e1eac92c890c3bb14299c97bf0f620e4ca...	897a167a0877(
08EEDF7B31D7BE5E363E546CE658AE75	e8520ea1290e75c600b99dc2e39580c5866184c03d14857c...	281219b846cdc
CA33DA5952D458380F3C1CE39A516CCF	03615fa8c6857b590fac3bfa746a37b14e07e862eae6821e42...	0e0f6c96ae7b6
67C60C675AEB04BCFBD5258C4714D8DB	43aa80b0020c67a58951bf8f0f2d45e1db77b2c33ea912e37d...	fd716024d4c73
1119F66F08E08F98C4ED73B24BAD8DD2	cc8d7a3f2c64c3262e7972300e26a6da7f686be2429b5f9025...	bde5f81e17640
AD98B9A7BB241332209A431C9C4CFD22	e12ab6b8d2e3b56f14008821db93b156bfb84c3fbd0909fdfbe...	c2fab8ecb0b21
CBEAB6D9ADA360CF7178EBDAA4C0749C	df6c5a8a9c0d0aa5312231549aea29a00b234ab86cb0c3aa4...	73d94ad40037
B8740C63D66BB43D7C82410A14FB7E10	117ee88b01be08edf5b2da6e96f3f69c229fd68ba7511cee6df...	3d04d9dc71eba

Governance Policies

- Identified data remains behind each hospital's firewall
- Credentialing at CWRU before access
- Common Data Use Agreement before accessing the data and/or creating a limited data set (LDS)
- Simplified IRB approval for LDS's
- IRB approval from each system for access to identified data (Reliant IRB)
- No comparative analyses across systems

Outline

- Introduction

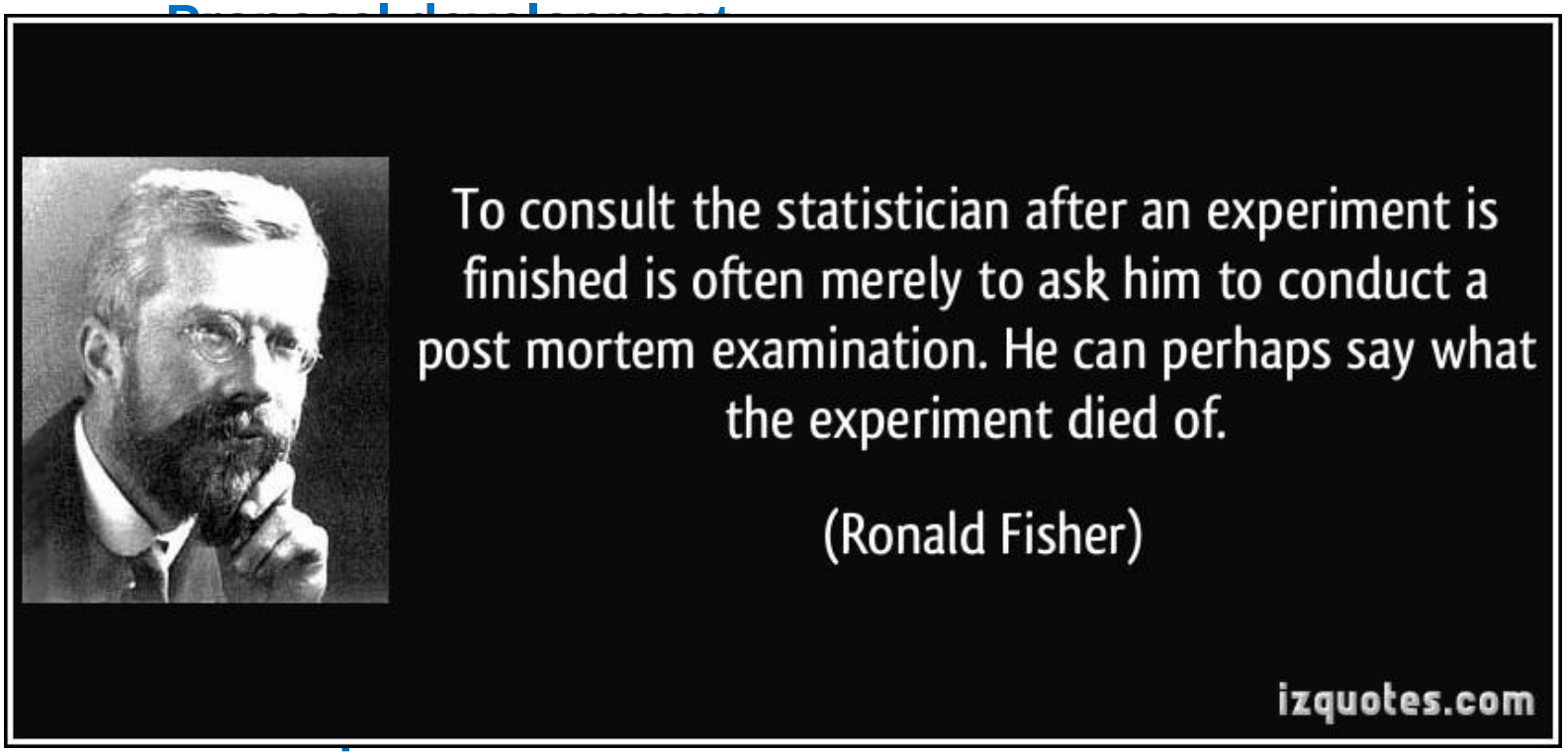
- Capturing the data

- **Analyzing the data**

- Connecting the data

EPBI Core Activities
Quantitative Cores

- **Study Design**



- **Interpreting the results appropriately**
- **Development of reports, presentations, and publications**

ICB Core Activities
Computational Biology Core

- **Provide analytical expertise for complex datasets**
- **Initial focus on genomics data**
  - **Processing and analysis of DNA sequences**
  - **Processing and analysis of RNA sequences**
  - **Processing and analysis of gene expression data**
  - ***In silico* functional annotation of variants**
- **Management to be integrated into the CWRU/CTSC Quantitative Cores umbrella**
- **Business model**
  - **Service model: chargebacks on a project/hourly basis**
  - **Collaborative model: salary and resource recovery directly via grants**

ICB Core Activities
Bioethics Core

- **Joint effort with Department of Bioethics**
- **Support development of policies and governance structures to address**
  - **Privacy**
  - **Informed consent (universal consent)**
  - **Data sharing**
- **Provide support for revising consent forms**
- **Provide support for interaction with the public**
  - **Community outreach**
  - **Educational materials**
- **Business Model**
  - **Collaborative activities**

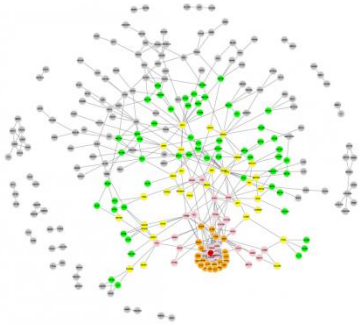
Outline

- Introduction

- Capturing the data

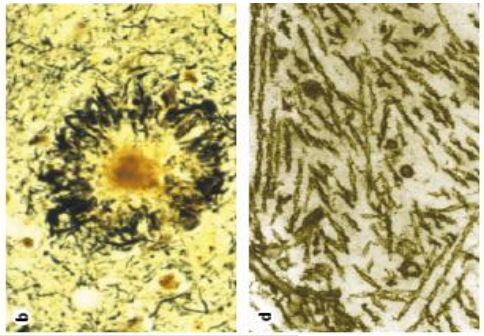
- Analyzing the data

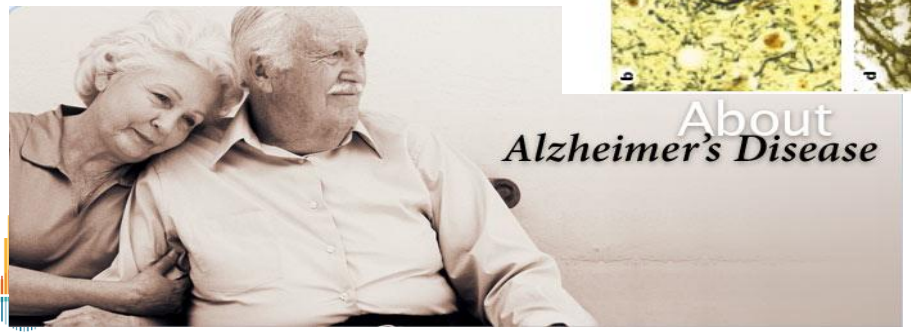
- **Connecting the data**

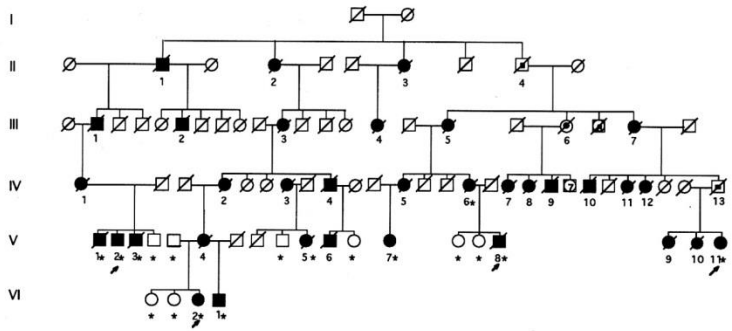


Alzheimer's Disease

- **Aging population.**
- **AD is common dementia with a shared pathology, but likely not etiology.**
- **No proven treatment or prevention (NIA mandate)**







Pedigree of FAD-RO1.

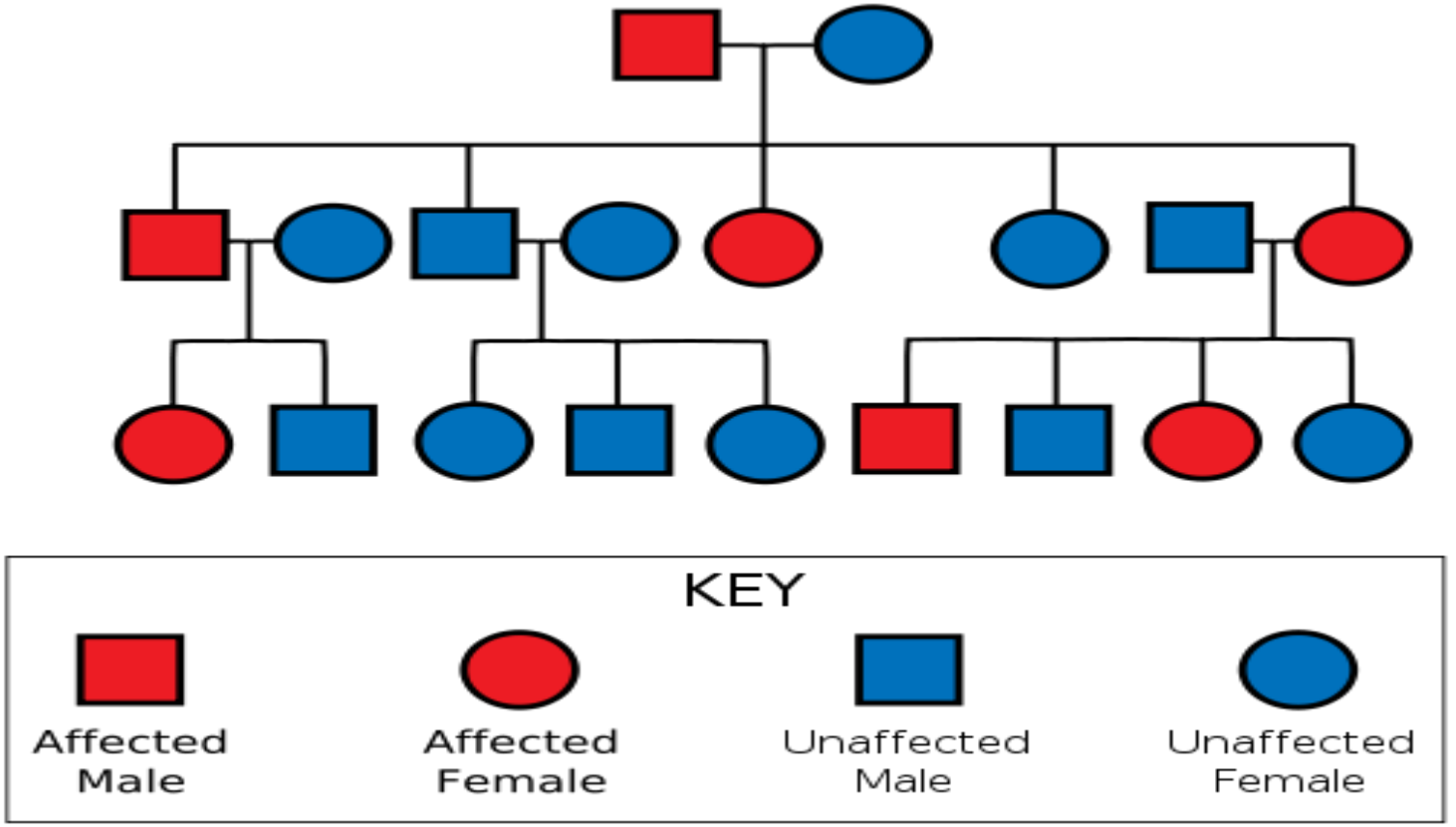
Integration of Alzheimer Research

Project	Acronym
Collaborative Aging and Memory Project	CAMP
The Alzheimer Disease Genetics Consortium	ADGC
The International Genetics of Alzheimer Project	IGAP
The Alzheimer Disease Sequencing Project	ADSP
Cleveland Alzheimer Disease Center	CADC

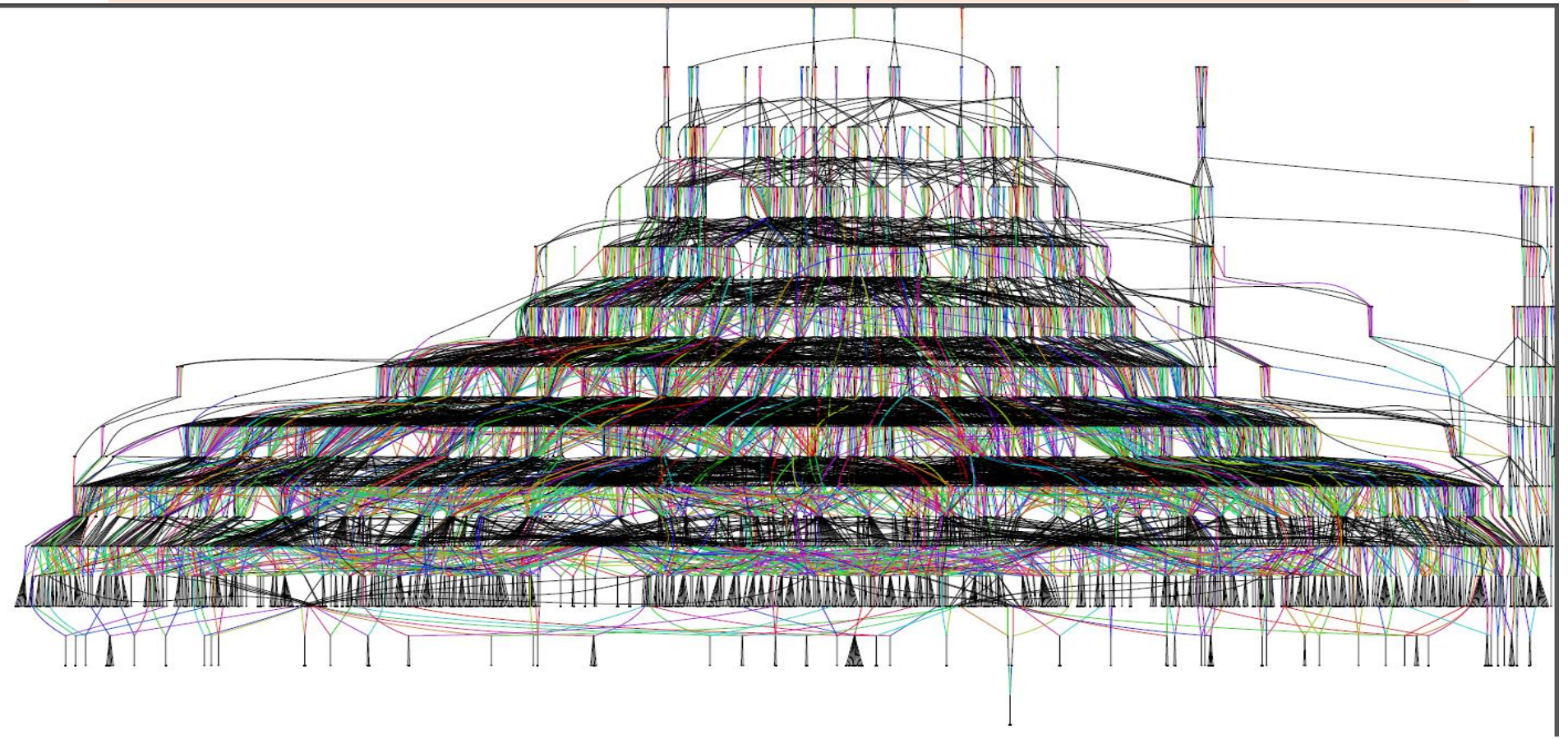
Collaborative Aging and Memory Project (CAMP)

- **Goal: Understand the genetics of successful aging and cognitive impairment**
- **Study population: Genetically isolated Amish communities in Ohio and Indiana**
- **Advantages:**
  - **Known Alzheimer genes are not as common in the Amish**
  - **Inter-related well documented pedigrees**
  - **Stable families**
  - **Relatively homogeneous environment**
- **Identified potentially important genetic targets**

Collaborative Aging and Memory Project (CAMP)

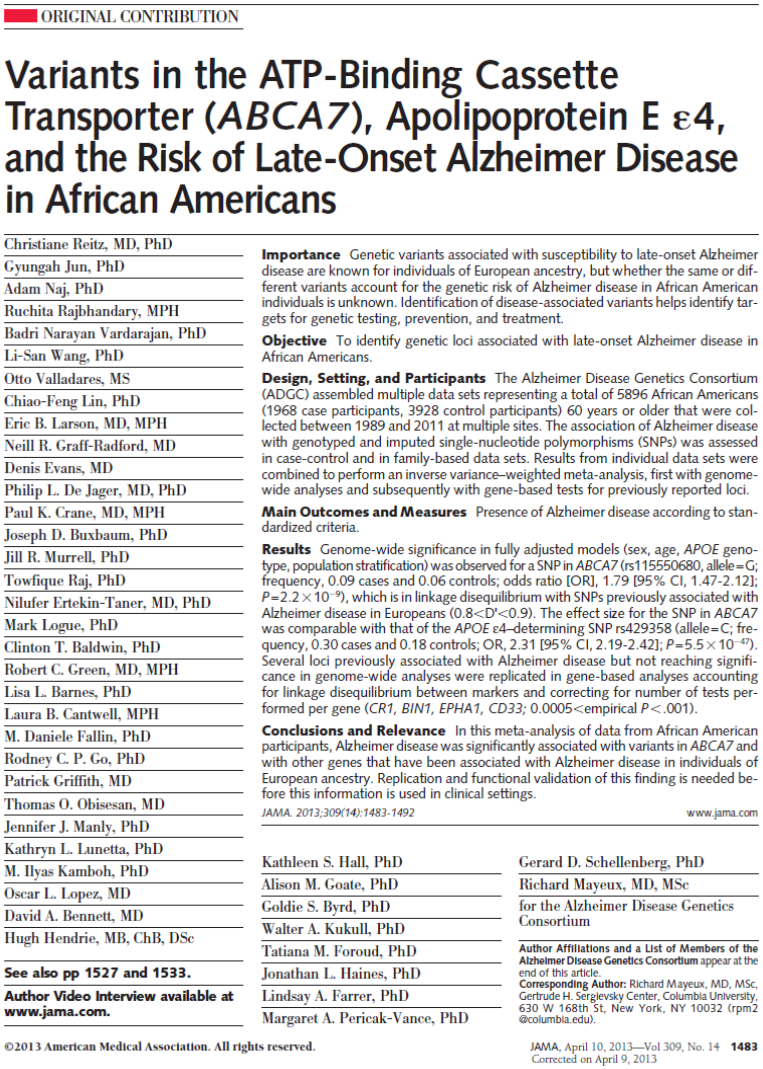


Collaborative Aging and Memory Project (CAMP)



Alzheimer Disease Genetics Consortium (ADGC)

- **Goal: Completely describe the genetics of Alzheimer disease**
- **Collaboration of > 15 U.S. based Alzheimer Genetics Researchers**
- **Samples of families, cases, and controls from over 20,000 cases and 15,000 controls**



ORIGINAL CONTRIBUTION

Variants in the ATP-Binding Cassette Transporter (*ABCA7*), Apolipoprotein E ε4, and the Risk of Late-Onset Alzheimer Disease in African Americans

Christiane Reitz, MD, PhD
Gyungah Jun, PhD
Adam Naj, PhD
Ruchita Rajbhandary, MPH
Badri Narayan Vardarajan, PhD
Li-San Wang, PhD
Otto Valladares, MS
Chiao-Feng Lin, PhD
Eric B. Larson, MD, MPH
Neill R. Graff-Radford, MD
Denis Evans, MD
Philip L. De Jager, MD, PhD
Paul K. Crane, MD, MPH
Joseph D. Buxbaum, PhD
Jill R. Murrell, PhD
Towfique Raj, PhD
Nilufer Ertekin-Taner, MD, PhD
Mark Logue, PhD
Clinton T. Baldwin, PhD
Robert C. Green, MD, MPH
Lisa L. Barnes, PhD
Laura B. Cantwell, MPH
M. Daniele Fallin, PhD
Rodney C. P. Go, PhD
Patrick Griffith, MD
Thomas O. Obisesan, MD
Jennifer J. Manly, PhD
Kathryn L. Lunetta, PhD
M. Ilyas Kamboh, PhD
Oscar L. Lopez, MD
David A. Bennett, MD
Hugh Hendrie, MB, ChB, DSc

See also pp 1527 and 1533.

Author Video Interview available at www.jama.com.

Importance Genetic variants associated with susceptibility to late-onset Alzheimer disease are known for individuals of European ancestry, but whether the same or different variants account for the genetic risk of Alzheimer disease in African American individuals is unknown. Identification of disease-associated variants helps identify targets for genetic testing, prevention, and treatment.

Objective To identify genetic loci associated with late-onset Alzheimer disease in African Americans.

Design, Setting, and Participants The Alzheimer Disease Genetics Consortium (ADGC) assembled multiple data sets representing a total of 5896 African Americans (1968 case participants, 3928 control participants) 60 years or older that were collected between 1989 and 2011 at multiple sites. The association of Alzheimer disease with genotyped and imputed single-nucleotide polymorphisms (SNPs) was assessed in case-control and in family-based data sets. Results from individual data sets were combined to perform an inverse variance–weighted meta-analysis, first with genome-wide analyses and subsequently with gene-based tests for previously reported loci.

Main Outcomes and Measures Presence of Alzheimer disease according to standardized criteria.

Results Genome-wide significance in fully adjusted models (sex, age, *APOE* genotype, population stratification) was observed for a SNP in *ABCA7* (rs115550680, allele=*G*; frequency, 0.09 cases and 0.06 controls; odds ratio [OR], 1.79 [95% CI, 1.47-2.12]; $P=2.2\times10^{-9}$), which is in linkage disequilibrium with SNPs previously associated with Alzheimer disease in Europeans ($0.8<D'<0.9$). The effect size for the SNP in *ABCA7* was comparable with that of the *APOE* ε4–determining SNP rs429358 (allele=C; frequency, 0.30 cases and 0.18 controls; OR, 2.31 [95% CI, 2.19-2.42]; $P=5.5\times10^{-47}$). Several loci previously associated with Alzheimer disease but not reaching significance in genome-wide analyses were replicated in gene-based analyses accounting for linkage disequilibrium between markers and correcting for number of tests performed per gene (*CR1, BIN1, EPHA1, CD33;* $0.0005<\text{empirical } P<.001$).

Conclusions and Relevance In this meta-analysis of data from African American participants, Alzheimer disease was significantly associated with variants in *ABCA7* and with other genes that have been associated with Alzheimer disease in individuals of European ancestry. Replication and functional validation of this finding is needed before this information is used in clinical settings.

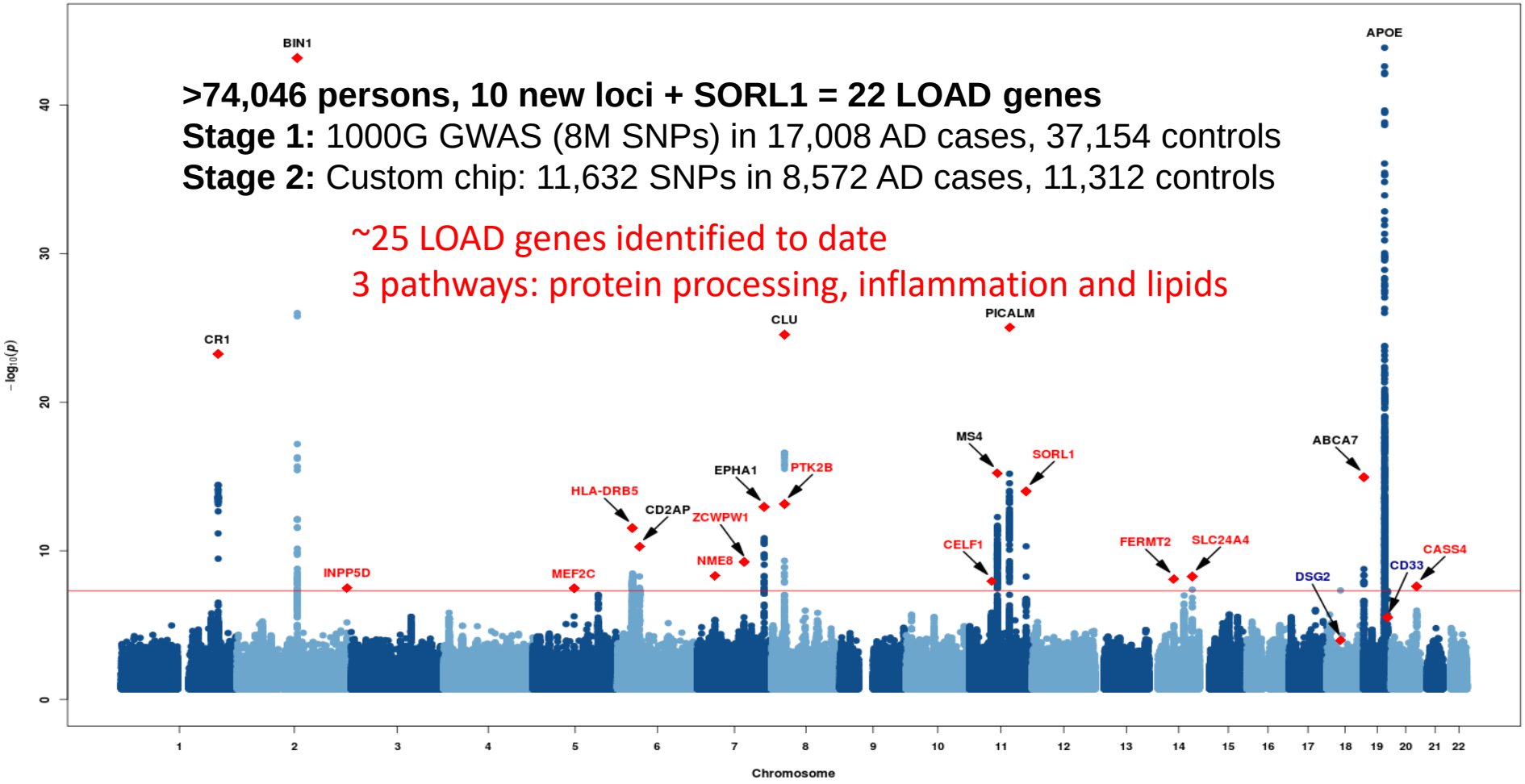
JAMA. 2013;309(14):1483-1492 www.jama.com

Kathleen S. Hall, PhD
Alison M. Goate, PhD
Goldie S. Byrd, PhD
Walter A. Kukull, PhD
Tatiana M. Foroud, PhD
Jonathan L. Haines, PhD
Lindsay A. Farrer, PhD
Margaret A. Pericak-Vance, PhD
Gerard D. Schellenberg, PhD
Richard Mayeux, MD, MSc
for the Alzheimer Disease Genetics Consortium

Author Affiliations and a List of Members of the Alzheimer Disease Genetics Consortium appear at the end of this article.
Corresponding Author: Richard Mayeux, MD, MSc, Gertrude H. Sergievsky Center, Columbia University, 630 W 168th St, New York, NY 10032 (rpm2@columbia.edu).

INSTITUTE FOR COMPUTATIONAL BIOLOGY

International Genetics of Alzheimer's Project (IGAP)



>74,046 persons, 10 new loci + SORL1 = 22 LOAD genes

Stage 1: 1000G GWAS (8M SNPs) in 17,008 AD cases, 37,154 controls

Stage 2: Custom chip: 11,632 SNPs in 8,572 AD cases, 11,312 controls

~25 LOAD genes identified to date

3 pathways: protein processing, inflammation and lipids

Alzheimer Disease Sequencing Project (ADSP)

- **February 7, 2012: Presidential Initiative announced to fight Alzheimer's Disease (AD)**
- **NIH to develop and execute a large scale sequencing project to understand AD genetics**
- **Immediate Goals:**
  - **Identify novel risk raising genetic variants**
  - **Identify novel protective genetic variants**
- **Long-term objective: to facilitate identification of new pathways for therapeutic approaches and prevention**

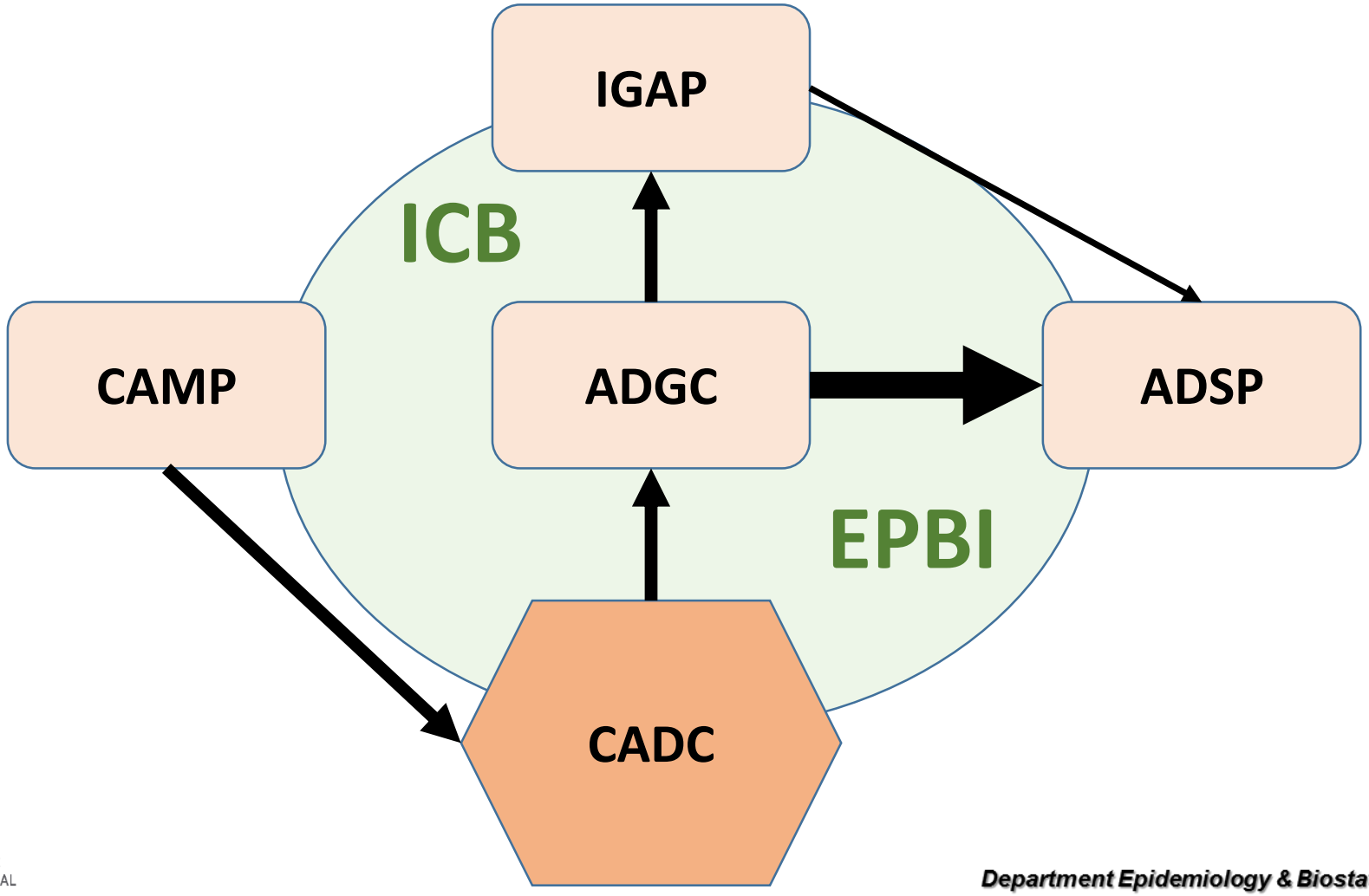
ADSP Dataset

- **Whole Genome DNA Sequencing**
  - **583 individuals in 111 families**
  - **6 billion bases/person**
  - **3.5 trillion bases of information**
- **Whole Exome DNA Sequencing**
  - **5,000 unrelated cases**
  - **1,000 enriched cases**
  - **5,000 elderly controls**
  - **75 million bases/person**
  - **825 billion bases of information**

Cleveland Alzheimer Disease Center (CADC) Initiative

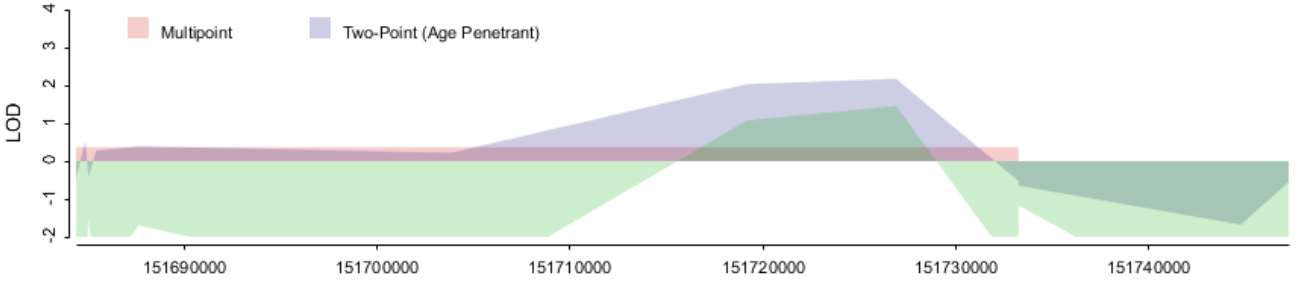
- **A collaborative effort**
  - **Cleveland Clinic (Leverenz)**
  - **University Hospitals (Lerner)**
  - **Case Western Reserve University (Haines)**
- **Alzheimer Disease Center goal is to support a wide range of Alzheimer disease research through**
  - **Clinical assessments**
  - **Research core support**
  - **Education**
  - **Outreach**

Alzheimer Disease Research Interactions

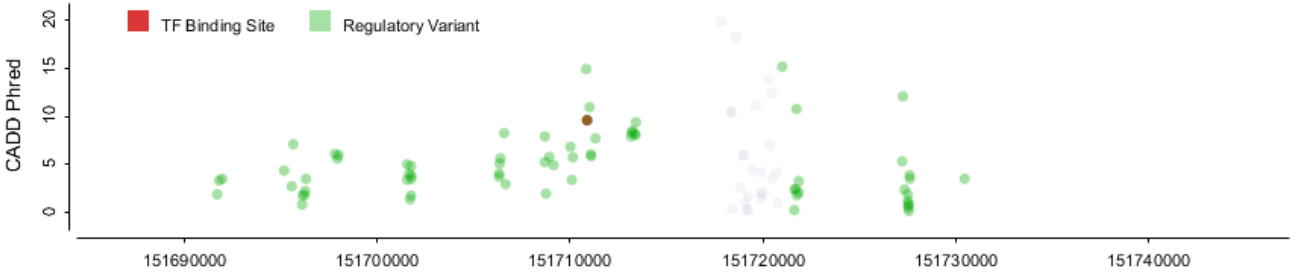


Linking Data Across Types

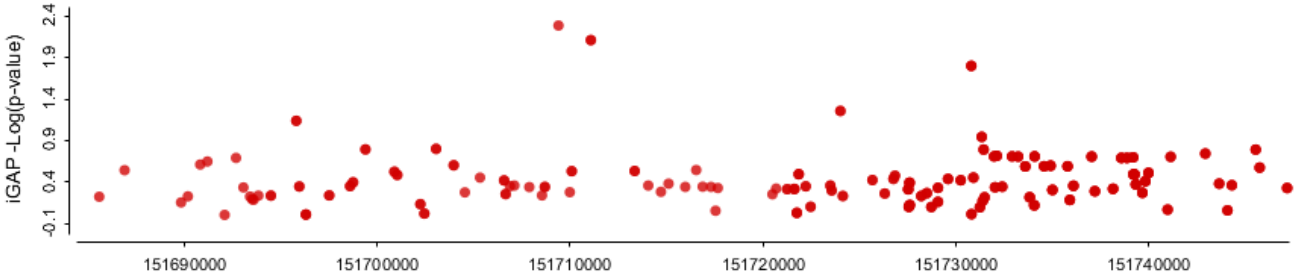
LOD
Chr6:151.6 – 151.7 MB



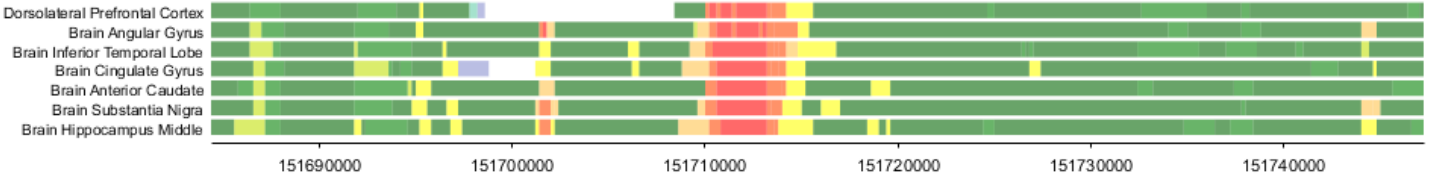
CADD Scores &
Regulatory Features



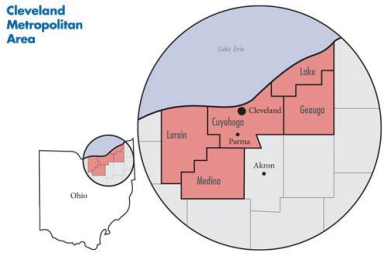
iGAP GWAS



Chromatin State
Segmentation



A Thought Experiment



- In 2014 University Hospitals had:
  - 923,000 unique patient visits
  - 84,000 surgeries
- Metropolitan Cleveland has 2,100,000 people
- Northeast Ohio has 4,300,000 people
- University Hospitals may have provided care for as many as
  - 40% of Metropolitan Cleveland
  - 20% of NEO
- If we collect data and samples on just 20% of the UH patients
  - 185,000 samples (e.g. blood)
  - 17,000 tissue samples
- What could you do with those data?

Integrating Data Types

Determinants of the Human Phenome