I2b2/tranSMART
BD2K PIC-SURE RESTful API

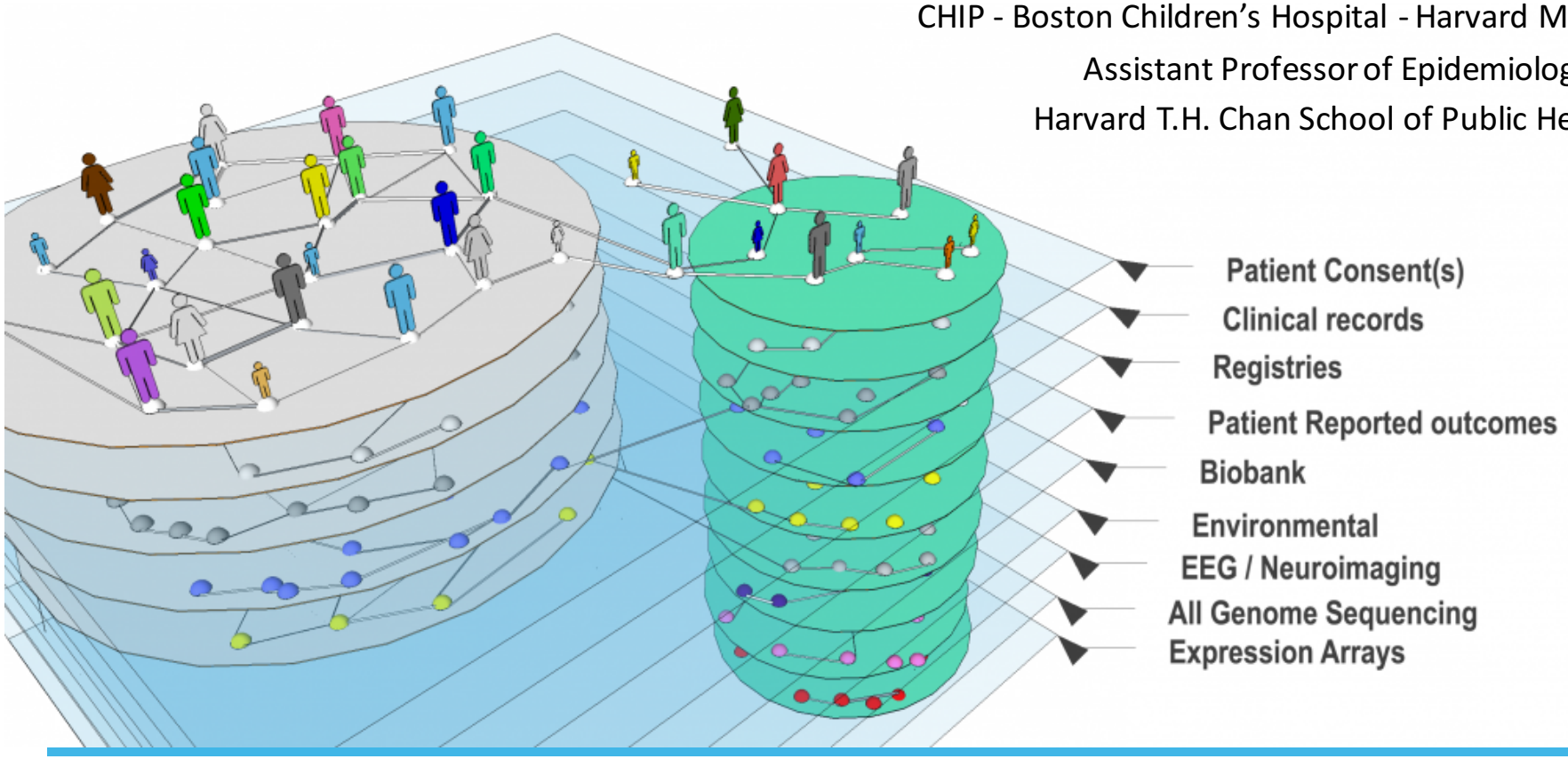
Paul Avillach, MD, PhD

Assistant Professor of Pediatrics and Biomedical Informatics

CHIP - Boston Children's Hospital - Harvard Medical School

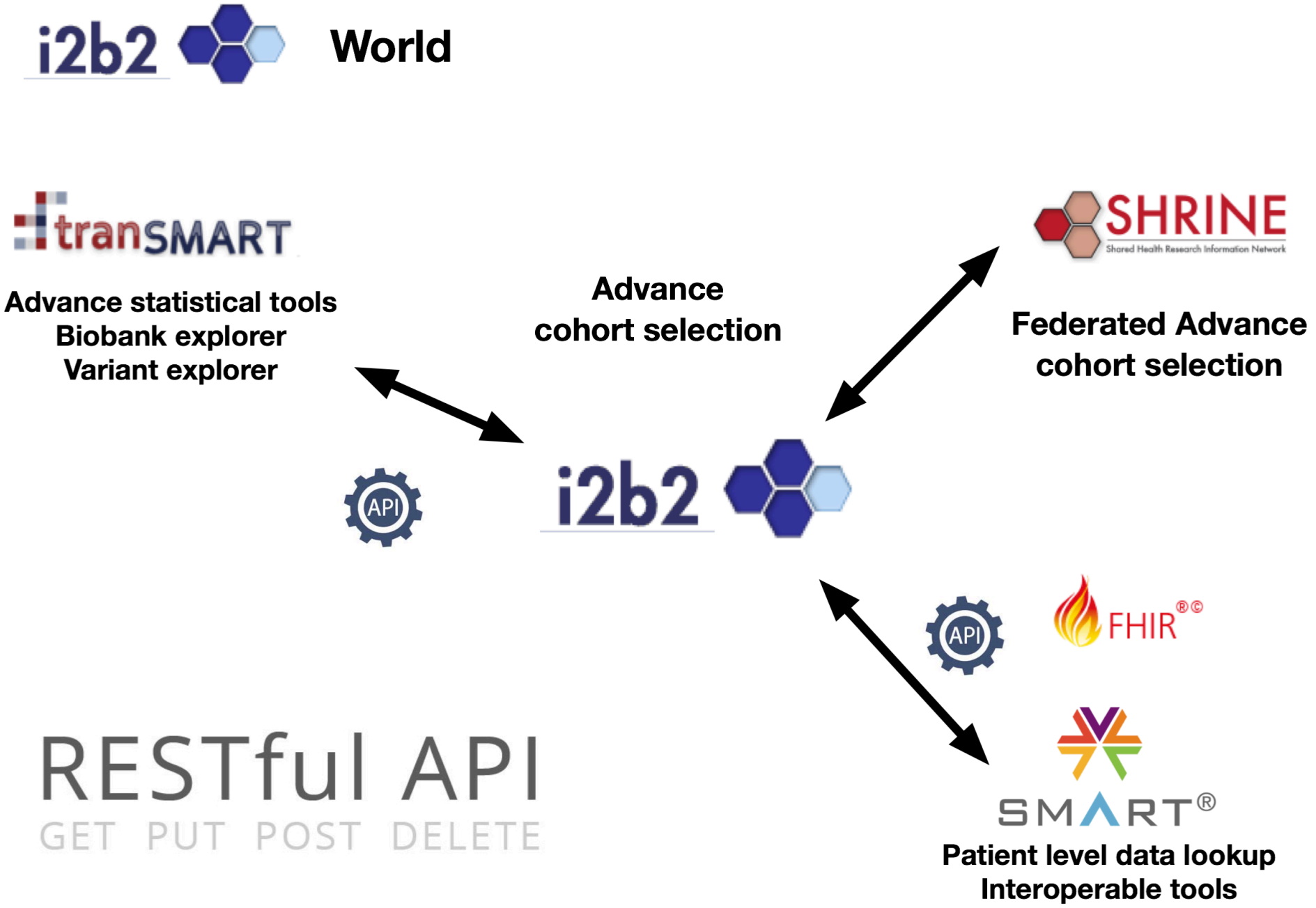
Assistant Professor of Epidemiology

Harvard T.H. Chan School of Public Health



HARVARD MEDICAL SCHOOL

DEPARTMENT OF Biomedical Informatics


i2b2
World
transMART
Advance statistical tools
Biobank explorer
Variant explorer
Advance
cohort selection
SHRINE
Shared Health Research Information Network
Federated Advance
cohort selection
API
i2b2
API
FHIR®©
RESTful API
GET PUT POST DELETE
SMART®
Patient level data lookup
Interoperable tools

CENTER FOR BIOMEDICAL INFORMATICS
HARVARD MEDICAL SCHOOL

Boston Children's Hospital Research Connection

EHR

i2b2

- Clinical notes
- ICD9
- Lab values
- Drugs

cTAKES

Research cohorts

Research cohorts

Research cohorts

Biobank Core

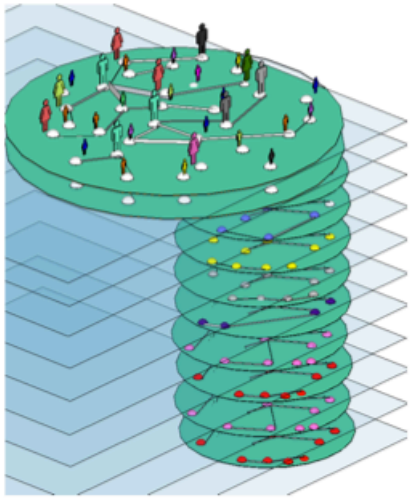
Patients' consent(s)

Selected patients

i2b2

tranSMART

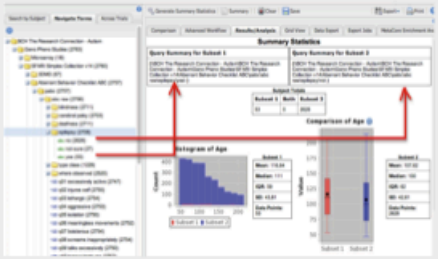
Patient centered Information Commons



'omics

Gene exp
SNPs
Whole Exome Seq
RNA seq

Generation of Hypotheses by PIs



"Drag & Drop"

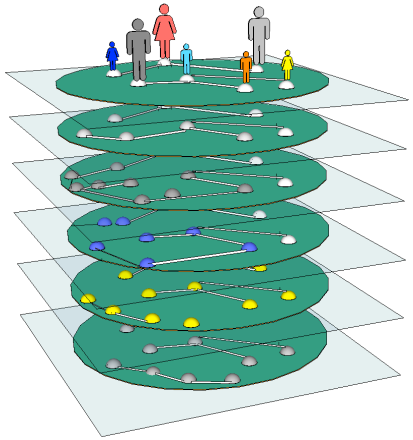
High throughput analysis by Biostatistician & Bioinformatician

BCH Private network

integration
analysis

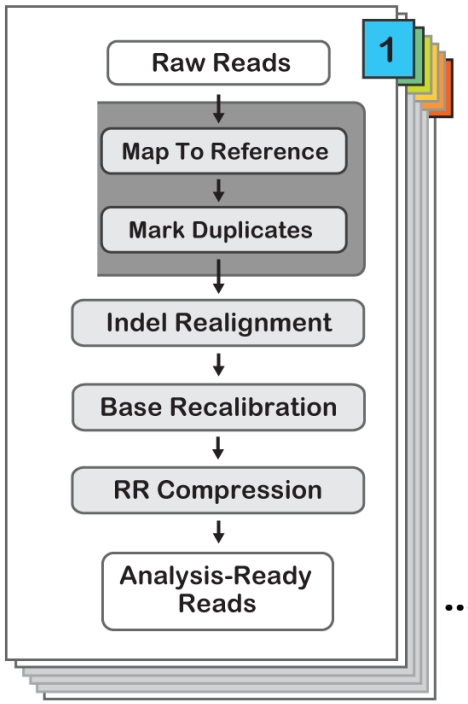
- Autism
  - BCH Consent Information (2975)
  - BCH EHR i2b2 (16587)
    - Allergy (8184)
    - Clinic (15547)
    - Demographics (16522)
    - Diagnoses (16426)
    - Insurance Payors (13949)
    - Labtests (12073)
    - Medications (9371)
    - Procedures (10140)
    - Radiology (8736)
    - Reports (4414)
    - Service (16587)
    - Vital Signs (11982)
  - Expression Array (713)
    - Peripheral Blood (713)
      - Affymetrix Human Genome U133 Plus 2.0 Array (198)
        - gcRMA Normalized Expression (198)
      - Human Gene 1.0 ST Array (550)
        - RMA Normalized Expression (550)
  - GenoPheno (3189)
  - SFARI Simplex Collection v14 (2760)
  - WES_Inventory (7373)
    - Eichler (5349)
    - State (2024)

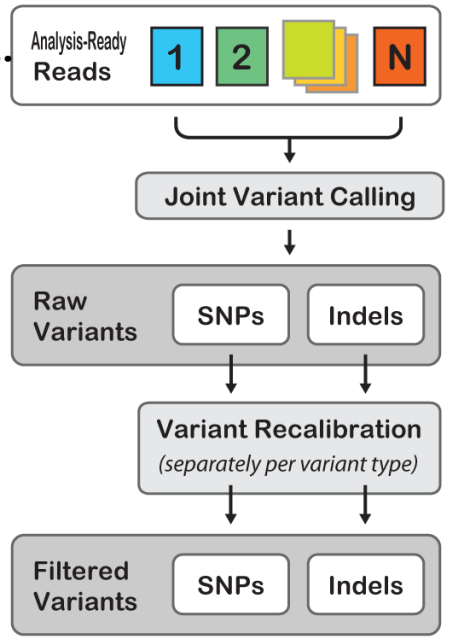
- **Patient consent(s)**
- **EHR longitudinal data**



- **Expression Arrays**
- **Clinical Cohorts**
- **WES data**

Exome sequence data processing > Variant calling > Variant annotation > I2b2/tranSMART input





gatk

Physical location
- e.g. Chr:start-end
- Cytoband
- …

Gene
- e.g. Gene name
- Variant function
- …

Gene set
- e.g. Pathway
- Molecular proces
- …

Predicted variant impact
- e.g. SIFT
- PolyPhen
- …

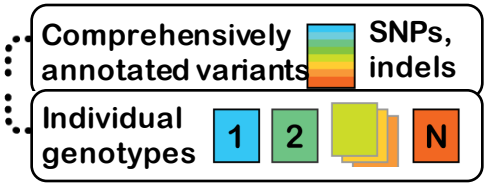
Conservation
- e.g. GERP
- PhyloP
- …

Population frequency
- e.g. 1000 Genomes
- ESP 6500
- …

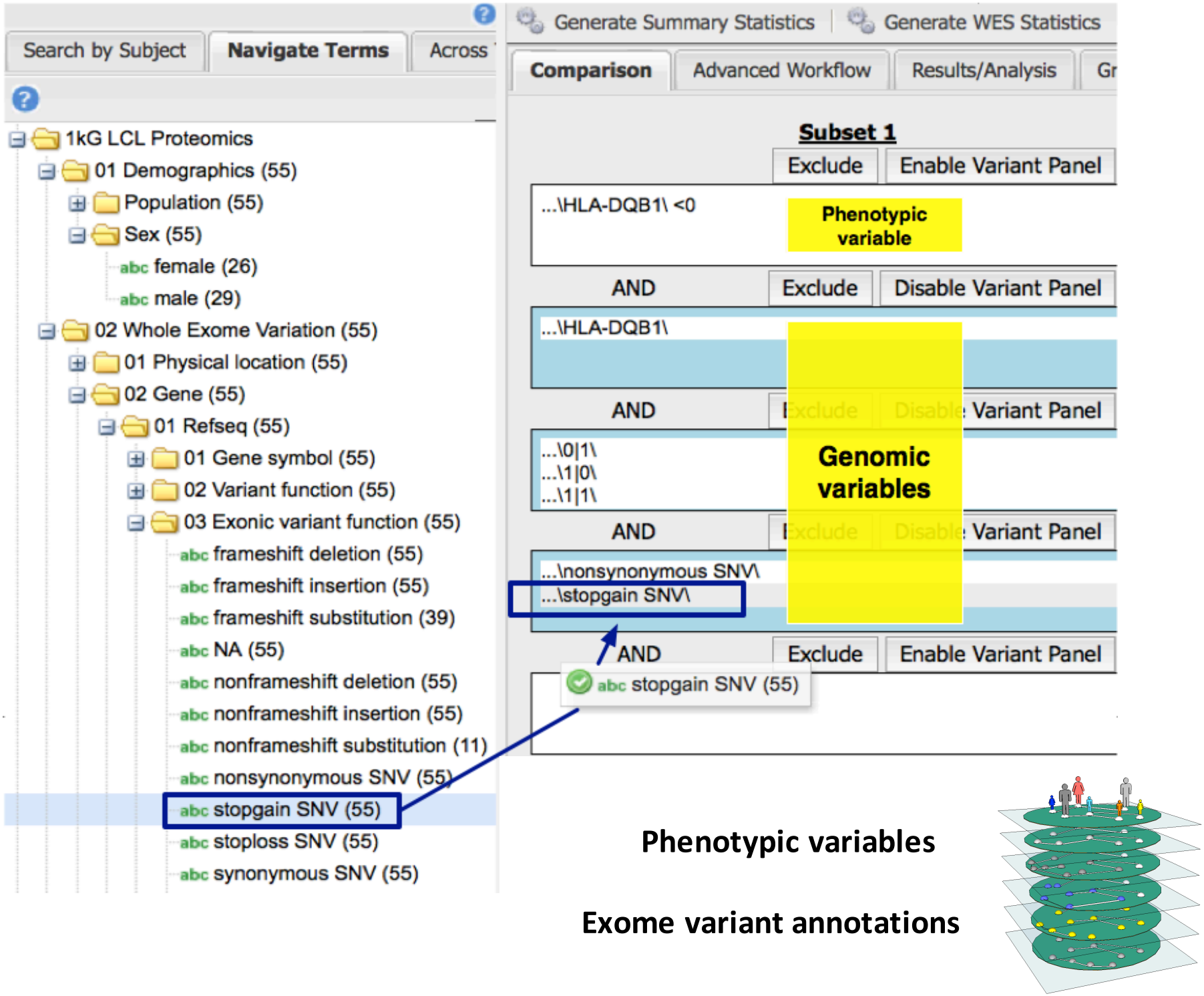
Clinical significance
- e.g. ClinVar
- OMIM
- …

Expression patterns
- e.g. GTEx
- BrainSpan
- …

Transcriptional regulatio
- e.g. ENCODE TFBS
- Histone modifications



ANNOVAR


Search by Subject
Navigate Terms
Across
1kG LCL Proteomics
01 Demographics (55)
Population (55)
Sex (55)
female (26)
male (29)
02 Whole Exome Variation (55)
01 Physical location (55)
02 Gene (55)
01 Refseq (55)
01 Gene symbol (55)
02 Variant function (55)
03 Exonic variant function (55)
frameshift deletion (55)
frameshift insertion (55)
frameshift substitution (39)
NA (55)
nonframeshift deletion (55)
nonframeshift insertion (55)
nonframeshift substitution (11)
nonsynonymous SNV (55)
stopgain SNV (55)
stoploss SNV (55)
synonymous SNV (55)
Generate Summary Statistics
Generate WES Statistics
Comparison
Advanced Workflow
Results/Analysis
Subset 1
Exclude
Enable Variant Panel
...\HLA-DQB1\ <0
Phenotypic variable
AND
Exclude
Disable Variant Panel
...\HLA-DQB1\
...\0|1\
...\1|0\
...\1|1\
Genomic variables
...\nonsynonymous SNV\
...\stopgain SNV\
stopgain SNV (55)
Phenotypic variables
Exome variant annotations

Harvard IRB security levels:

Level 5 - Extremely sensitive information

Level 4 – Very sensitive information

Level 3 – Sensitive, or Confidential information

Level 2 - Benign information to be held confidentially

Level 1 - Non-confidential research information

HARVARD MEDICAL SCHOOL

HIPAA COMPLIANT

AWS

VPC

PIC-SURE

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

CCD - PIC-SURE

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

CountEverything

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

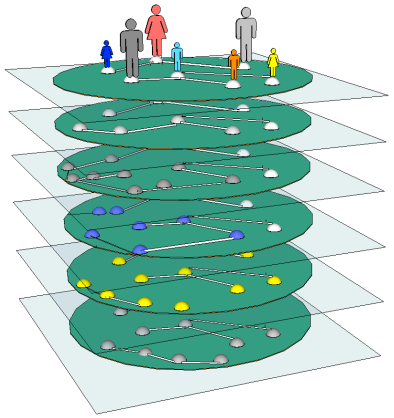
GRDR

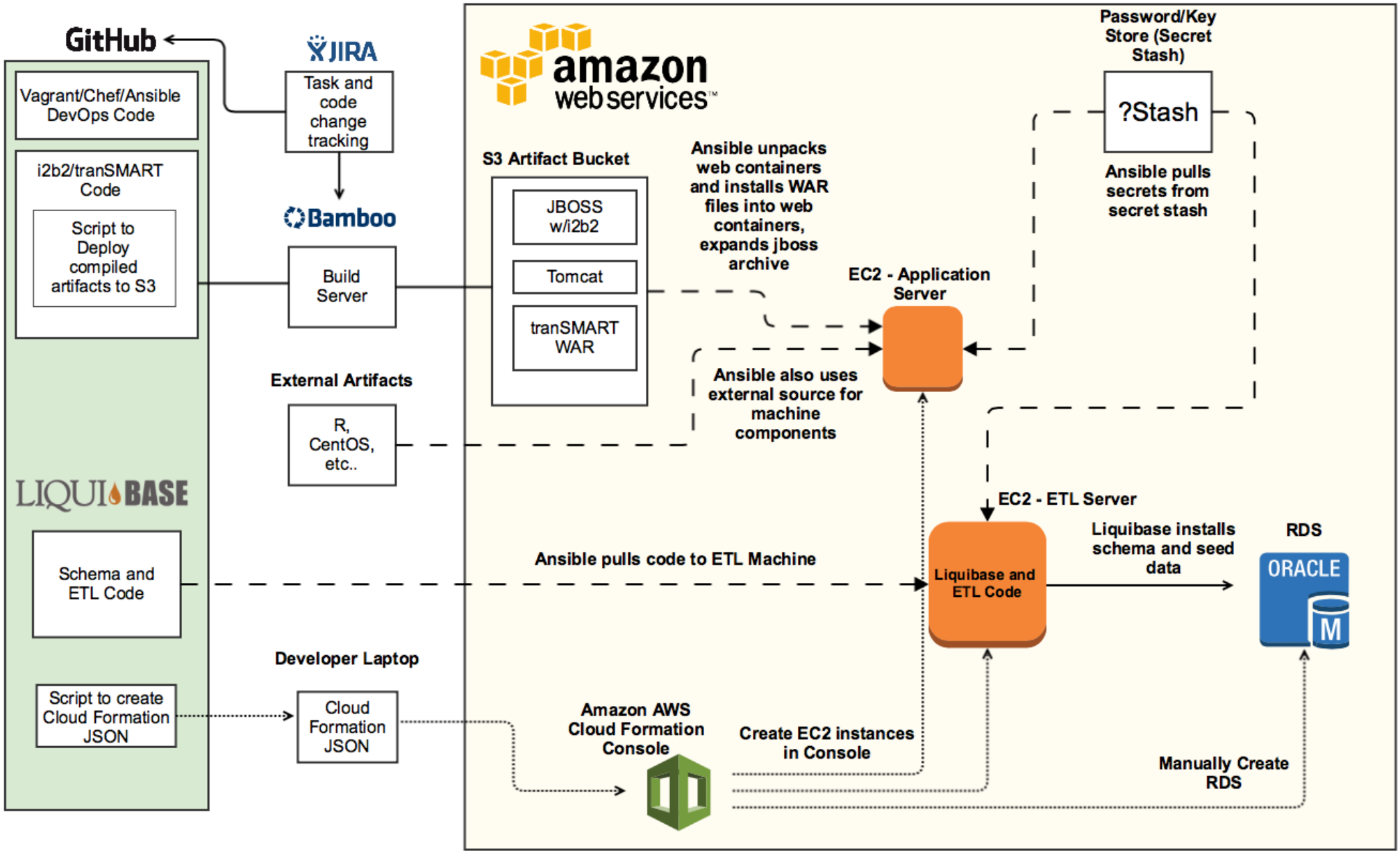
NIH National Center for Advancing Translational Sciences

VPC

pcori **PPRN PMS_DN**

.....




GitHub
Vagrant/Chef/Ansible DevOps Code
i2b2/tranSMART Code
Script to Deploy compiled artifacts to S3
LIQUIBASE
Schema and ETL Code
Script to create Cloud Formation JSON
JIRA
Task and code change tracking
Bamboo
Build Server
External Artifacts
R, CentOS, etc..
Developer Laptop
Cloud Formation JSON
amazon web services
S3 Artifact Bucket
JBOSS w/i2b2
Tomcat
tranSMART WAR
Ansible unpacks web containers and installs WAR files into web containers, expands jboss archive
Ansible also uses external source for machine components
Password/Key Store (Secret Stash)
?Stash
Ansible pulls secrets from secret stash
EC2 - Application Server
EC2 - ETL Server
Ansible pulls code to ETL Machine
Liquibase and ETL Code
Liquibase installs schema and seed data
RDS
ORACLE
Amazon AWS Cloud Formation Console
Create EC2 instances in Console
Manually Create RDS

HIPPA Compliance on AWS

Secured access control

1. **Authentication**

   Who are you?

2. **Authorization**

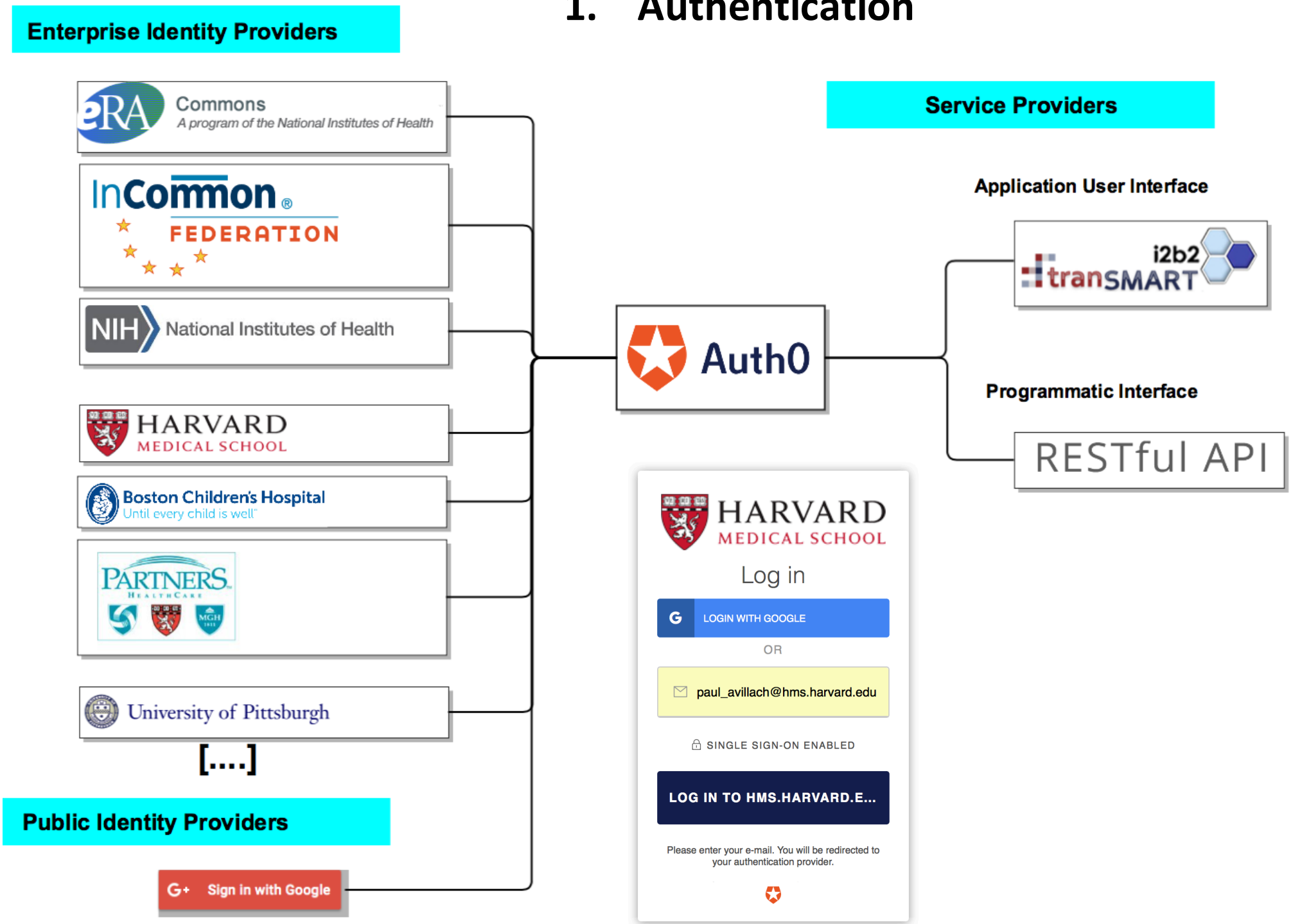
   What are you allowed to do?

   Level 0 *: Authenticated BUT no access to data*

   Level 1 *: Aggregated data*

   Level 2 ***:*** *patient level data*

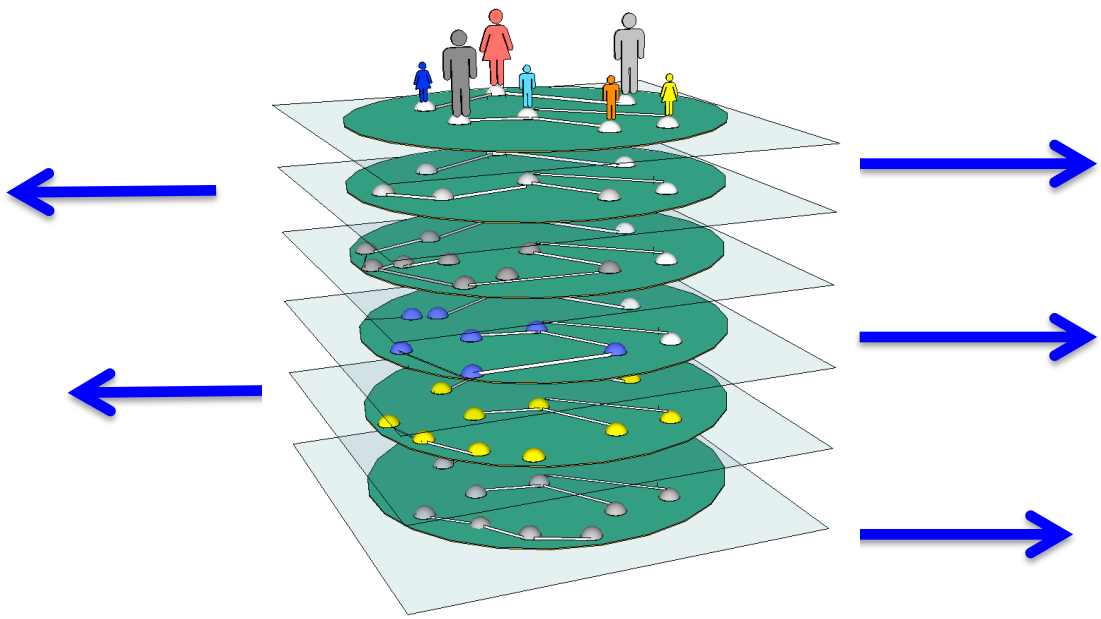
1. Authentication



NIH Big Data to Knowledge (BD2K)

Patient Centric Information Commons (PIC)

amazon web services™



HIPAA
Health Insurance Portability and Accountability Act

RESTful API

Play with API: https://bd2k-picsure.hms.harvard.edu

Source code: https://github.com/hms-dbmi


NIH
NIH Big Data to
Knowledge (BD2K)

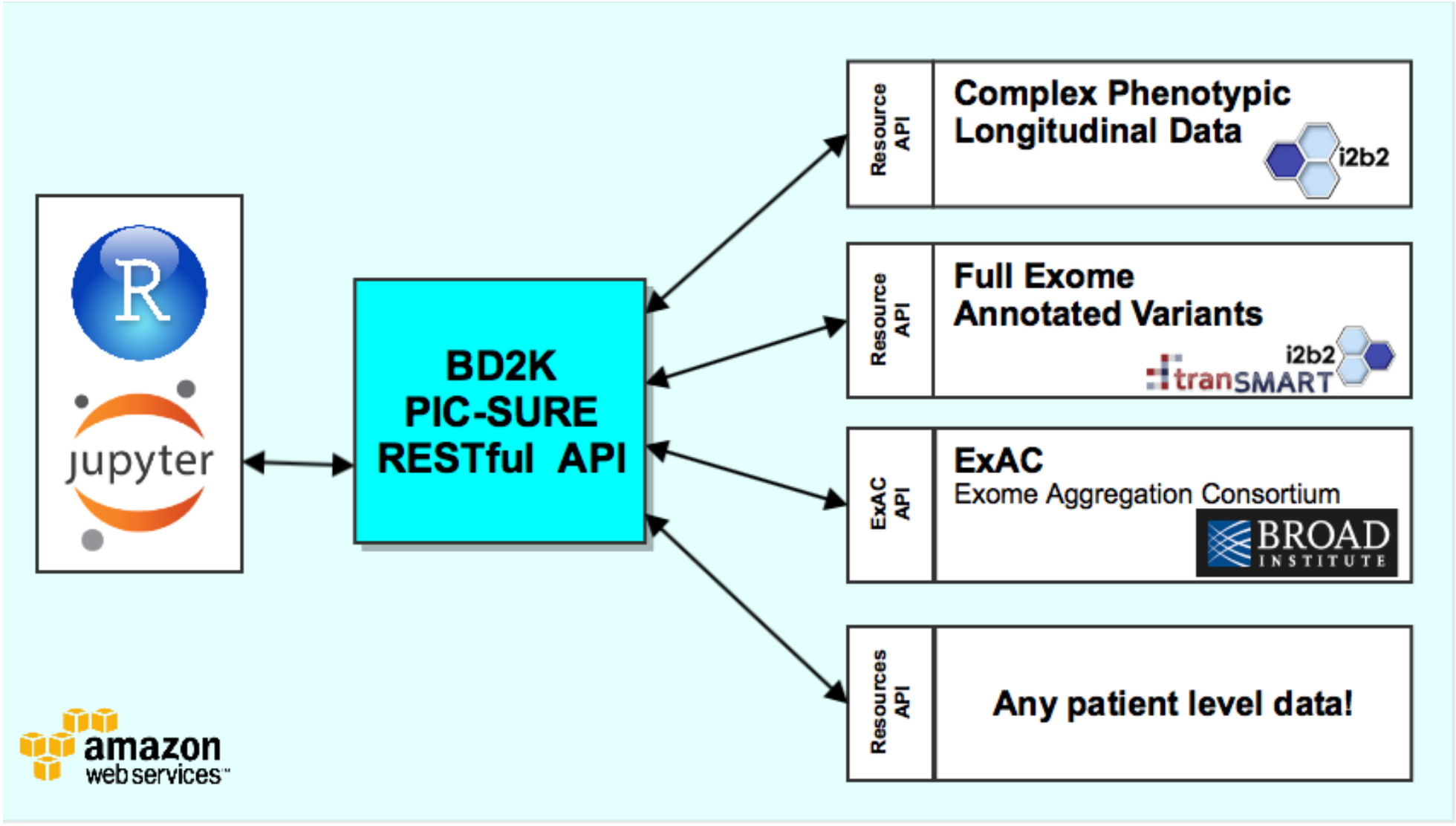

PIC-SURE


amazon
web services™


HIPAA
Health Insurance Portability
and Accountability Act


i2b2

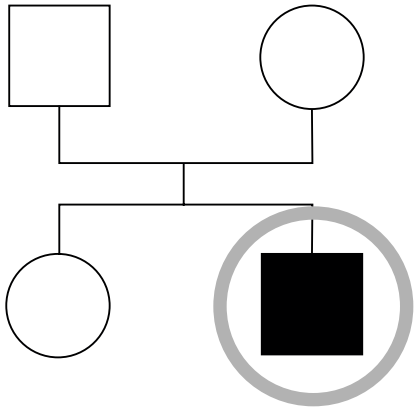

HARVARD
MEDICAL SCHOOL
DEPARTMENT OF
Biomedical Informatics



Play with PIC-SURE API: https://bd2k-picsure.hms.harvard.edu

Play with ExAC API: http://exac.hms.harvard.edu

Simons Simplex Collection: phenotypic data



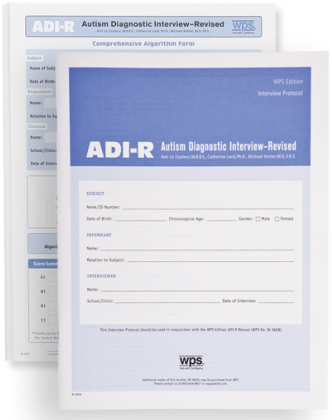
2,800+ quad simplex families

27 standardized questionnaires:

- Neurodevelopmental assessment tools
- Background

Quantitative, categorical and binary variables

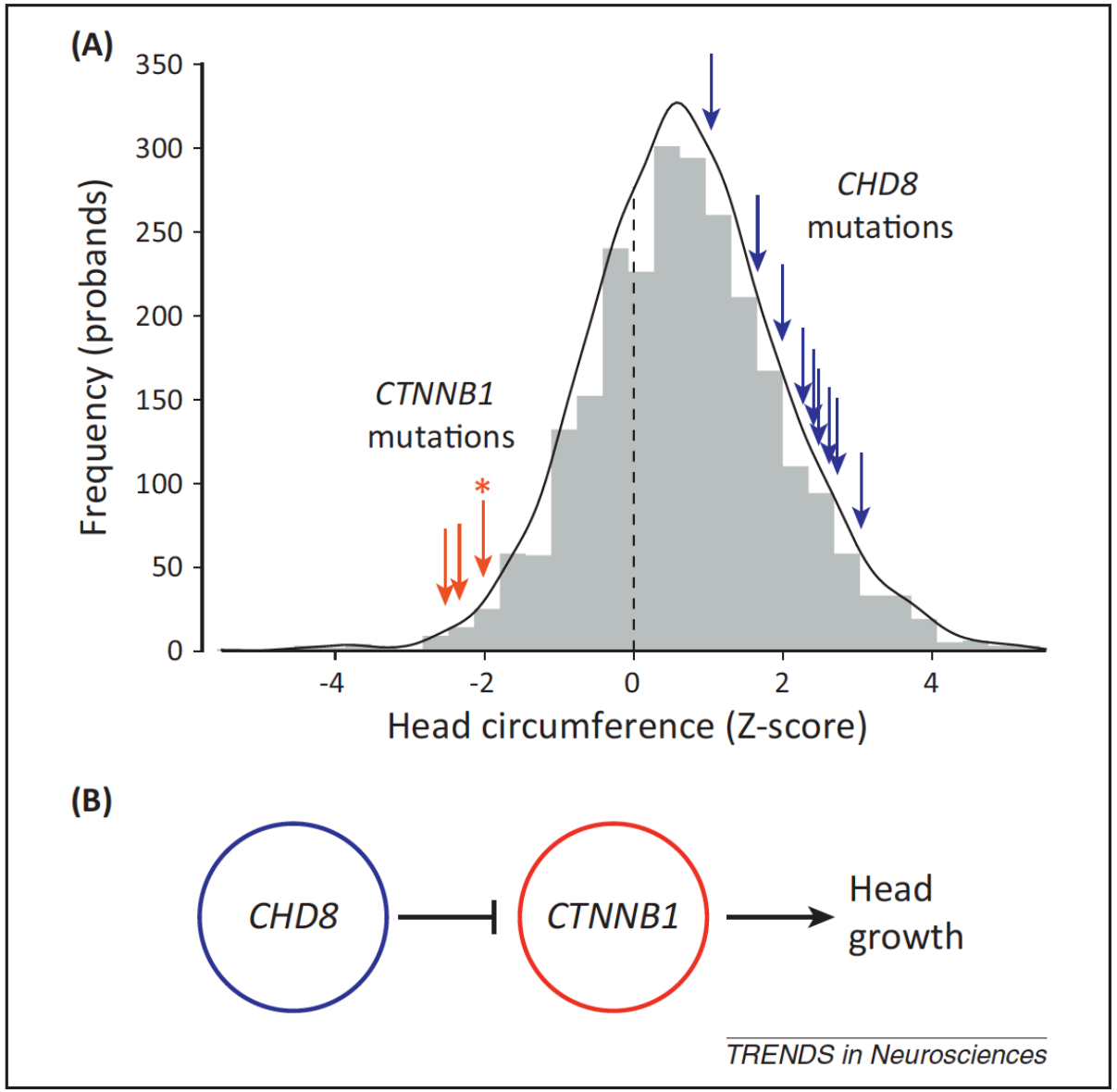
6,000 clinical variables



SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Simons Simplex collection

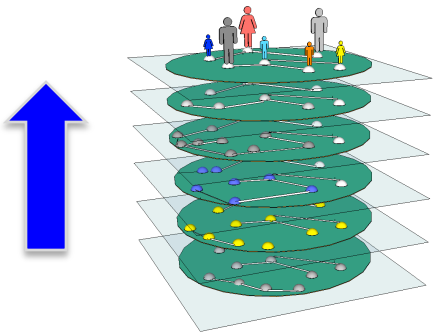
ABC	RBS-R
ABCL	SSC-Background
ADI-R	SSC-Diagnosis
ADOS	SSC-HWHC
BAPQ	SSC-MHI
CBCL	SSC-Pedigree
CTOPP-NR	SSC-THF
C-TRF	SQC
DAS-II	SRS
DCDQ	SRS-ARV
FHI-I	TRF
Mullen	WISC-IV
Purdue Pegboard	Vineland-II
	WASI



CHD8 mutated =>
Larger Head circumference

Phenotypes

Genotypes

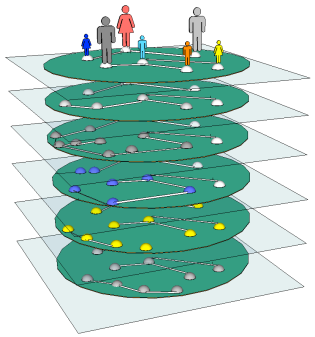


Krumm N, O'Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. 2014 Feb;37(2):95-105.

Live demo

Phenotypes

Genotypes



HARVARD MEDICAL SCHOOL

HIPAA COMPLIANT

AWS

VPC

PIC-SURE

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

CCD - PIC-SURE

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

CountEverything

NIH *NIH Big Data to Knowledge (BD2K)*

VPC

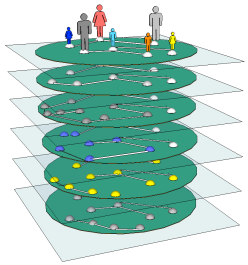
GRDR

NIH National Center for Advancing Translational Sciences

VPC

pcori **PPRN PMS_DN**

.....



Boston Children's Hospital

Until every child is well™

HIPAA COMPLIANT

AWS

VPC

Precision Link Biobank portal

.....

PARTNERS HEALTHCARE

MGH MASSACHUSETTS GENERAL HOSPITAL

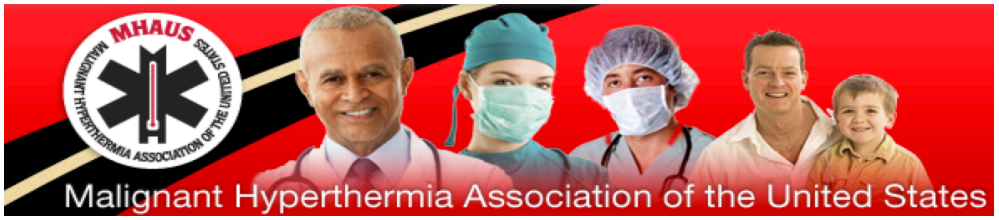
NGRID

NIH National Institute of Mental Health

MVP

National Center for Advancing Translational Sciences

The NIH/NCATS GRDR® Program
Global Rare Diseases Patient Registry
Data Repository



Marshfield Clinic
Research Foundation

Clinical Registry Investigating Bardet-Biedl Syndrome (CRIBBS)

INTRACRANIAL HYPERTENSION
RESEARCH FOUNDATION
CONTACT

National Ataxia Foundation

IWSA

INTERNATIONAL WAGR SYNDROME ASSOCIATION

Pachyonychia Congenita Project
Fighting for a cure. Connecting & helping patients. Empowering research.

The Plastic Bronchitis Foundation
Looking for a Cause, Working on a cure, Education and assisting

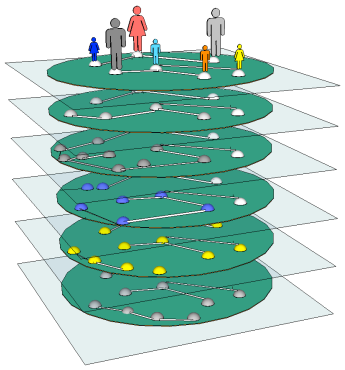
PBF

Wolfram Syndrome International Registry and Clinical Study

CdLS Foundation
Cornelia de Lange Syndrome Foundation, Inc.

https://grdr.hms.harvard.edu

Rare disease registry	Patients	Variables
Clinical Registry Investigating Bardet-Biedl Syndrome	180	708
International Pachyonychia Congenita Research Registry	569	496
International Plastic Bronchitis registry	66	63
Intracranial Hypertension Registry	1,349	91
North American Malignant Hyperthermia Registry	2,122	154
Wolfram Syndrome International Registry	124	580
Coordination of Rare Diseases at Sanford Registry	2,091	40
including:	*including:*	*Additional variables:*
National Ataxia Foundation	*869*	*15*
International WAGR syndrome association	*52*	*379*
Cornelia De Lange Syndrome Registry	*67*	*485*
Total	**6,501**	**2,132**



https://grdr.hms.harvard.edu

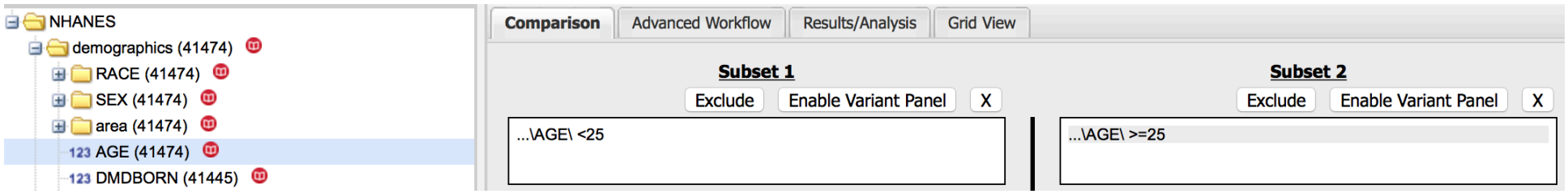
CDC National Health and Nutrition Examination Survey

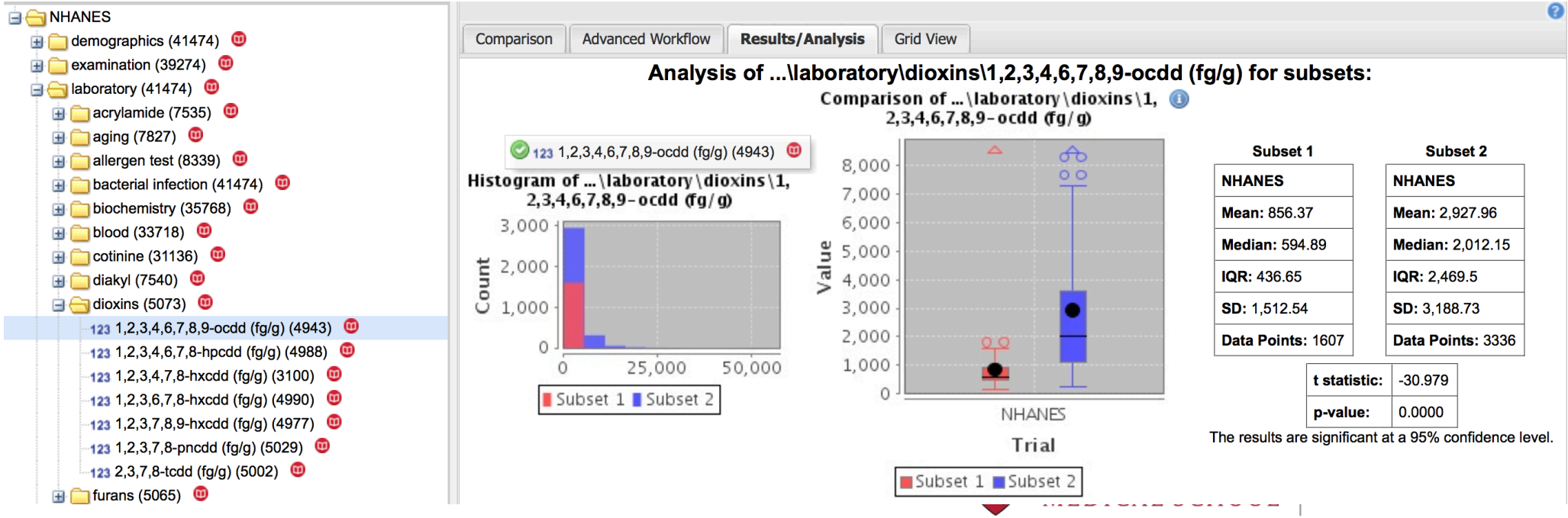
41k patients

2k patient level environmental variables

I2b2/tranSMART user Interface: https://nhanes.hms.harvard.edu

login: demo pass: demo

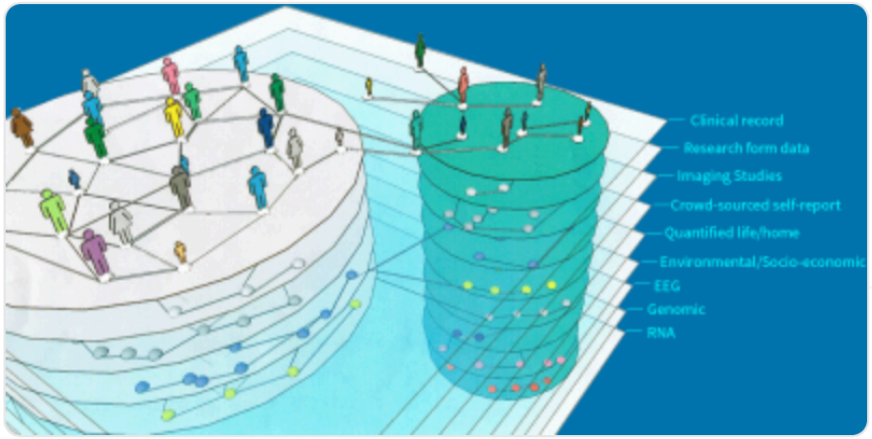




PIC-SURE Patient-centered Information Commons: Standardized Unification of Research Elements

A BD2K Center of Excellence ★

About the Project | About Us | Blog and News | Events | Training | Software | Publications | Contact Us



Creating an information commons for biomedical data centered on the patient.

PIC-SURE combines genetic, environmental, imaging, behavioral, and clinical data on individual patients from multiple sources into integrated sets.

learn more

www.pic-sure.org

- **Mentors**
  - Isaac Kohane (DBMI)
  - Susanne Churchill (DBMI)

- **Collaborators**
  - Ken Mandl (BCH)
  - David Margulies (BCH)
  - Sek Won Kong (BCH)
  - Joel Hirschhorn (BCH)
  - Florence Bourgeois (BCH)
  - Jon Bickel (BCH)
  - Ally Eran (BCH)
  - Tim Yu (BCH)
  - Savova Guergana (BCH)
  - Doug McFaden (HMS)
  - Shawn Murphy (Partners)
  - Chirag Patel (DBMI)
  - Nathan Palmer (DBMI)
  - David Bernick (Broad)
  - Finale Doshi-Velez (MIT)
  - Peter Szolovits (MIT)
  - Simon Lovestone (Oxford, UK)
  - Michelle Williams (Harvard T Chan SPH)
  - Tianxi Cai (Harvard T Chan SPH)
  - Greg Cooper (U Pitt.)
  - Lucila Ohno-Machado (UCSD)
  - David Haussler (UCSD)
  - Ida Sim (UCSF)
  - Anthony J Brookes (U. Leicester, UK)
  - Johan van Der Lei (Erasmus MC, NL)
  - Anita Burgun (Paris, FR)
  - Riccardo Bellazzi (U. Pavia, IT)
  - Mehan O'Bolye (PMS foundation)
  - Geraldine Bliss (PMS foundation)
  - Thomas Bourgeron (Institut Pasteur, FR)

- **DBMI Administrative support**
  - Katherine Flannery
  - Sunny Alvear
  - Jennifer Grandfield
  - + all DBMI staff

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

Avillach Lab

http://avillach-lab.hms.harvard.edu

Manager of Data Infrastructure

- **Michael McDuffie, MSc**

Developers

- Jeremy Easton-Marks
- Gabor Korodi
- Thomas DeSain
- Sean Finan
- Ranjay Kumar
- Alexander Nikitin
- Ken Hoflen

Graduate students

- Ombeline Dorval, MD
- Maxime Wack, MD
- Claire Hassen-Kodja, MD, MSc
- Emmanuelle Sylvestre, MD, MSc
- Yuri Ahuja, MD, PhD HMS candidate

Project Manager

Cassandra Perry, MS, CGC

Research Associates - Postdocs

- Antoine Tran, MD, MSc
- Laurie Tran, MD, MSc
- Cartik Saravana, PhD
- Li Ly, PhD
- Joany Zachariasse, MD, MSc
- Antoine Neuraz, MD, MSc

Previous members

- Samuel Finlayson, MD, PhD HMS candidate
- Ephi Sachs, MD
- Pei Chen
- Sushma Hanawal

Funding:

NIH NIH Big Data to Knowledge (BD2K)
NIH National Center for Advancing Translational Sciences
amazon web services
THE NANCY LURIE MARKS FAMILY FOUNDATION
pcori Patient-Centered Outcomes Research Institute

We are hiring now:

- Senior Software Developer *2
- Postdocs *3