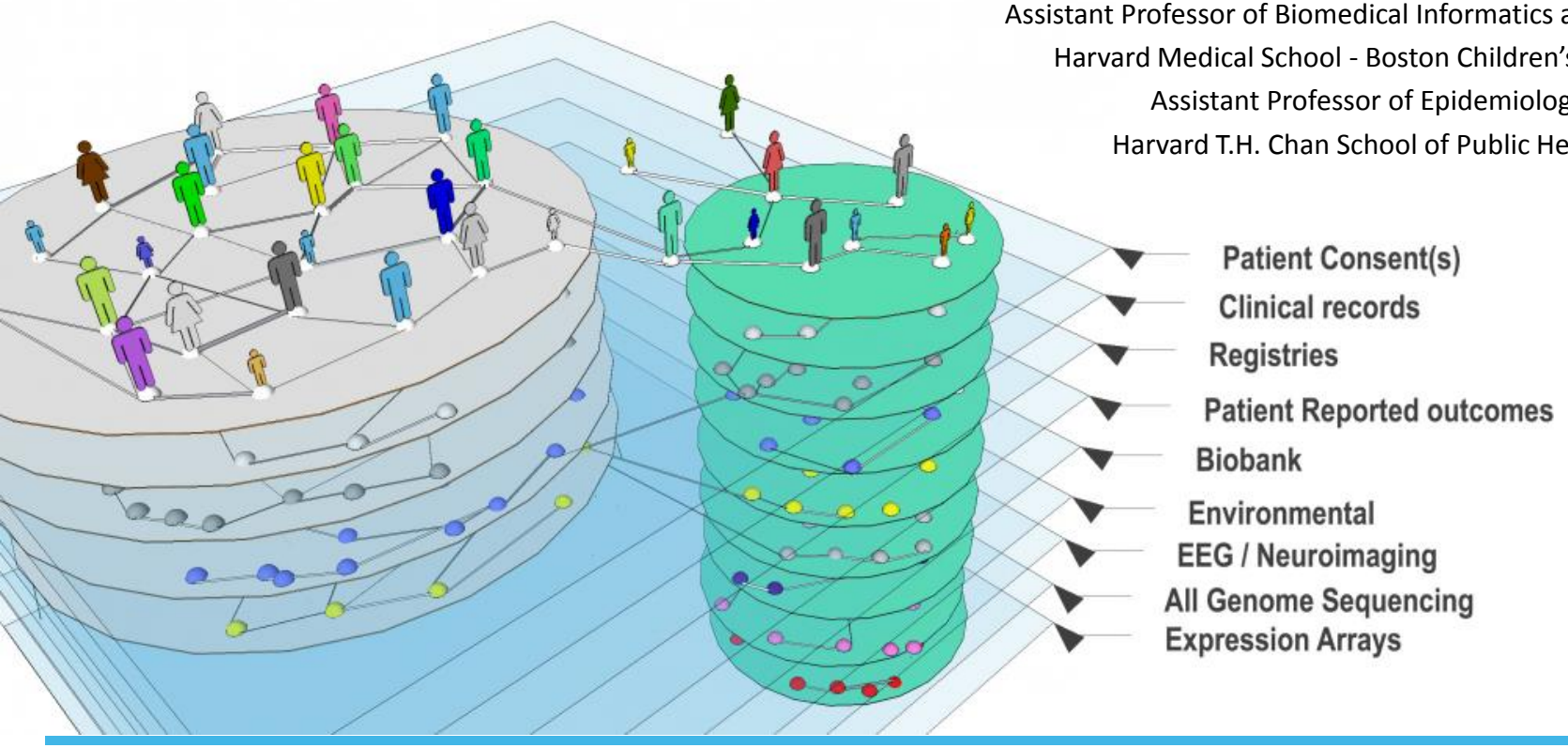
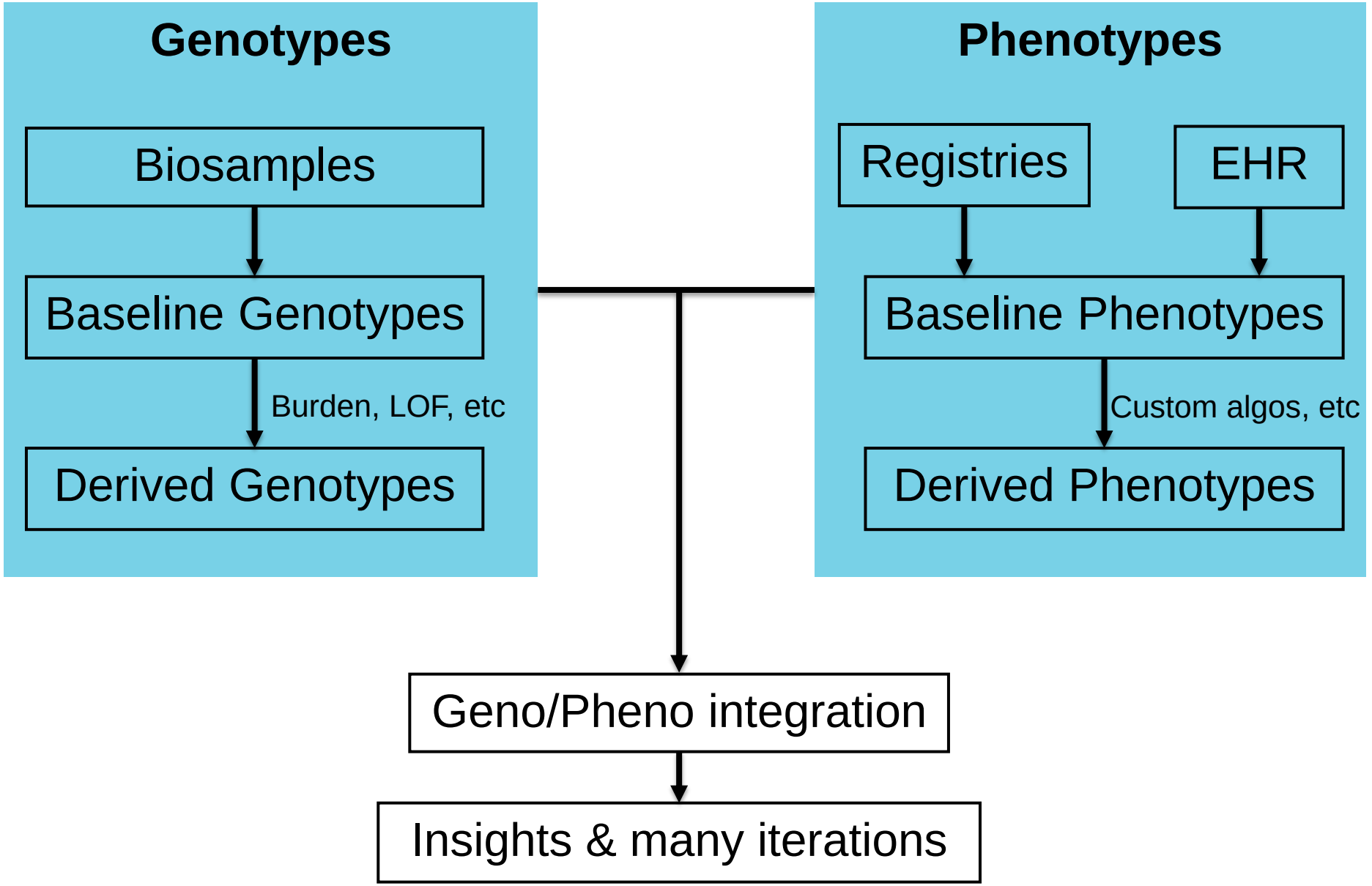
Creating scalable, secured, clinical and genomics platforms across clouds integrating i2b2 and tranSMART Platforms

Paul Avillach, MD, PhD
Assistant Professor of Biomedical Informatics and Pediatrics
Harvard Medical School - Boston Children's Hospital
Assistant Professor of Epidemiology
Harvard T.H. Chan School of Public Health



HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics


Genotypes
Biosamples
Baseline Genotypes
Burden, LOF, etc
Derived Genotypes
Phenotypes
Registries
EHR
Baseline Phenotypes
Custom algos, etc
Derived Phenotypes
Geno/Pheno integration
Insights & many iterations

World

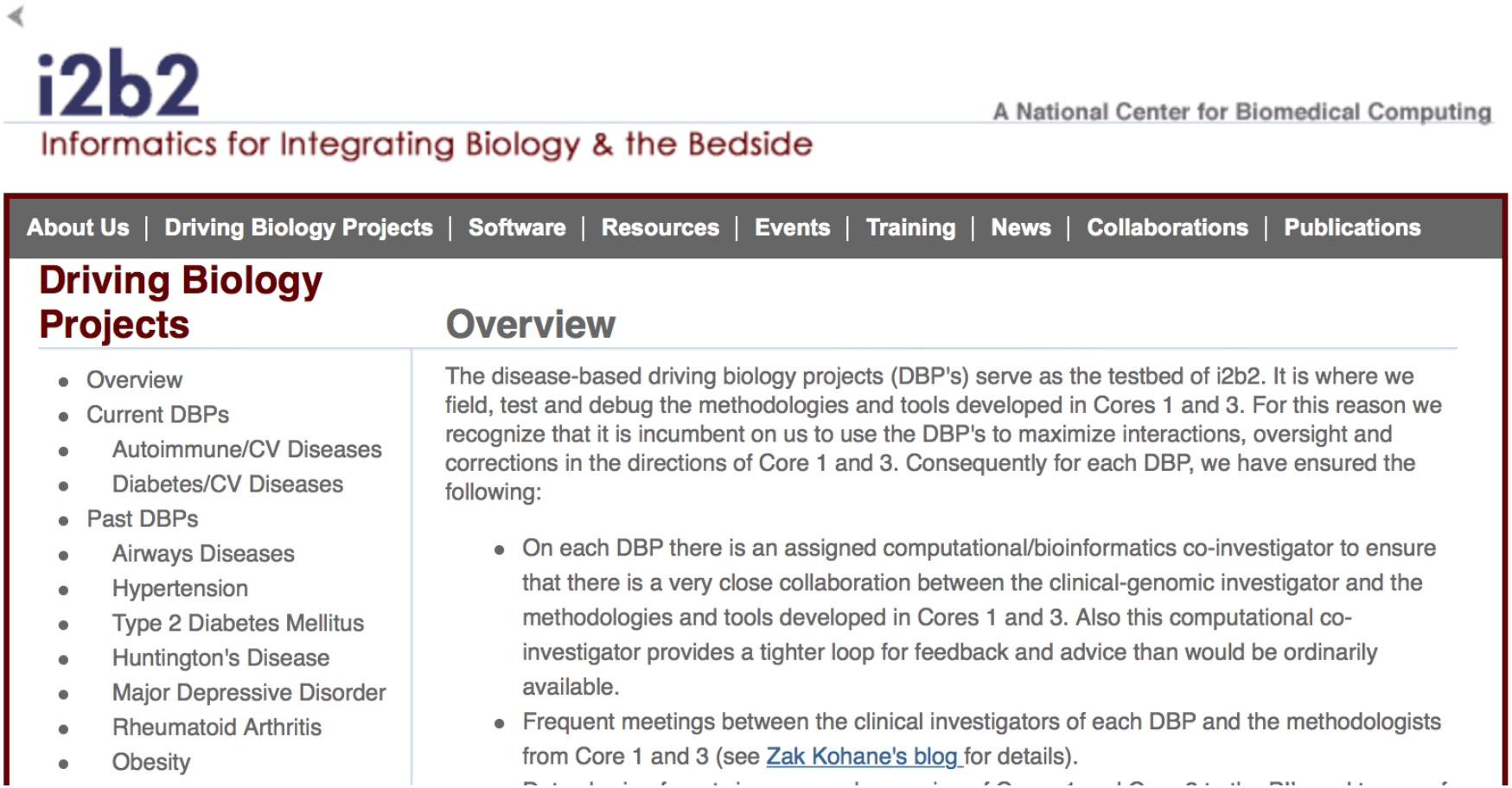
Advance statistical tools
Biobank explorer
Variant explorer

Advance
cohort selection

Federated Advance
cohort selection

Patient level data lookup
Interoperable tools

https://www.i2b2.org/disease/index.html



i2b2

Informatics for Integrating Biology & the Bedside

A National Center for Biomedical Computing

About Us | Driving Biology Projects | Software | Resources | Events | Training | News | Collaborations | Publications

Driving Biology Projects

- Overview
- Current DBPs
  - Autoimmune/CV Diseases
  - Diabetes/CV Diseases
- Past DBPs
  - Airways Diseases
  - Hypertension
  - Type 2 Diabetes Mellitus
  - Huntington's Disease
  - Major Depressive Disorder
  - Rheumatoid Arthritis
  - Obesity

Overview

The disease-based driving biology projects (DBP's) serve as the testbed of i2b2. It is where we field, test and debug the methodologies and tools developed in Cores 1 and 3. For this reason we recognize that it is incumbent on us to use the DBP's to maximize interactions, oversight and corrections in the directions of Core 1 and 3. Consequently for each DBP, we have ensured the following:

- On each DBP there is an assigned computational/bioinformatics co-investigator to ensure that there is a very close collaboration between the clinical-genomic investigator and the methodologies and tools developed in Cores 1 and 3. Also this computational co-investigator provides a tighter loop for feedback and advice than would be ordinarily available.
- Frequent meetings between the clinical investigators of each DBP and the methodologists from Core 1 and 3 (see Zak Kohane's blog for details).

eTRIKS

European Translational Information and Knowledge Management Services

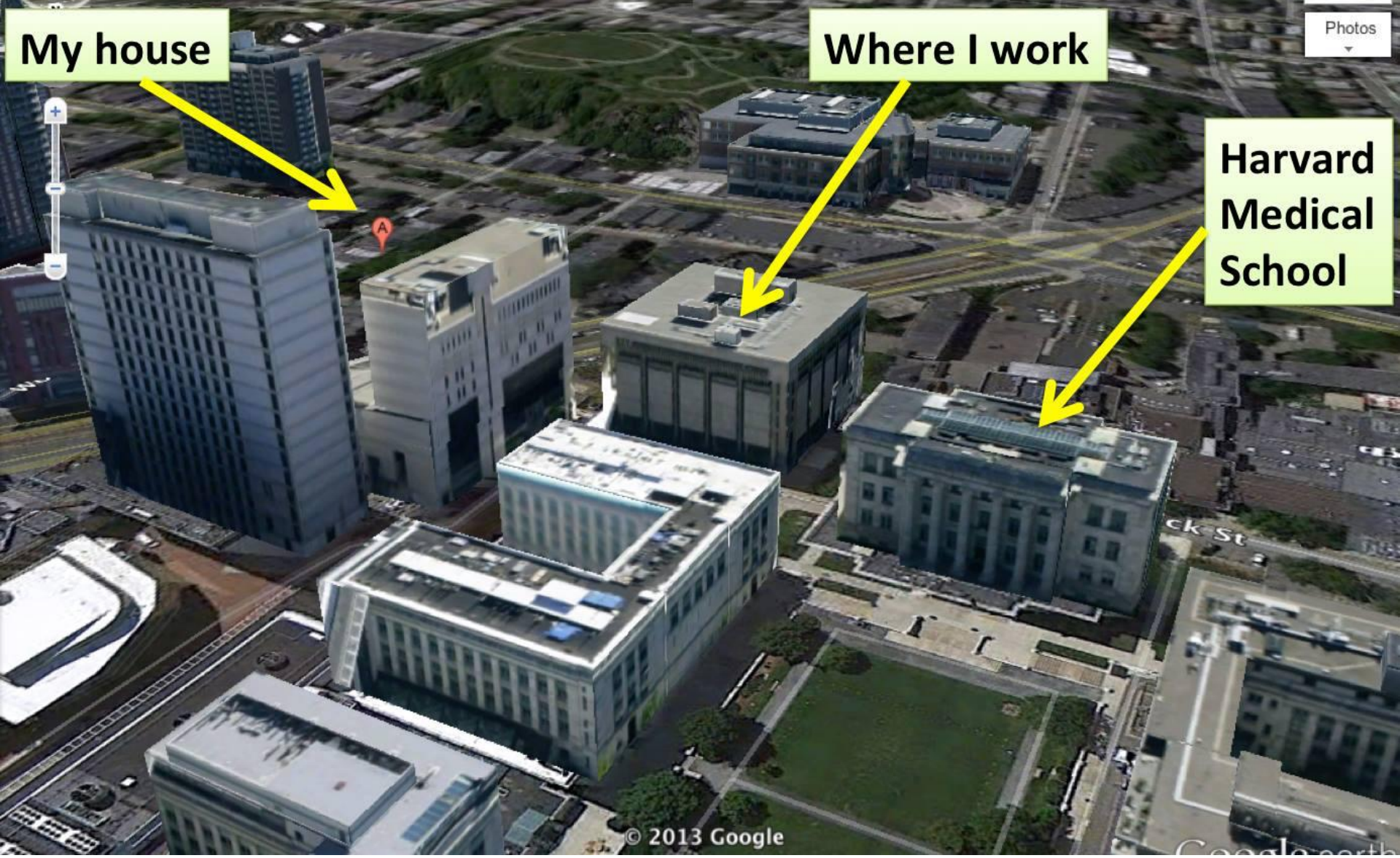
efpia

imi

tranSMART

v1.1

I2b2 database still there
But NOT i2b2 jboss app


My house
Where I work
Harvard Medical School
Photos
© 2013 Google

CENTER FOR BIOMEDICAL INFORMATICS
HARVARD MEDICAL SCHOOL

Boston Children's Hospital Research Connection

EHR

i2b2

- Clinical notes
- ICD9
- Lab values
- Drugs

cTAKES

Selected patients

Research cohorts

Research cohorts

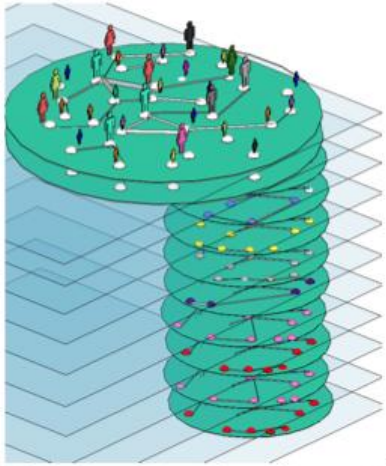
Research cohorts

Biobank Core

Patients' consent(s)

i2b2 tranSMART

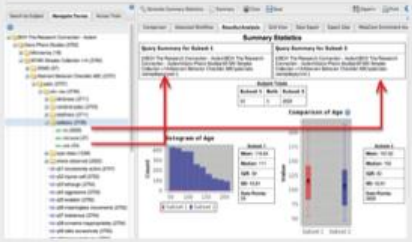
Patient centered Information Commons



'omics

Gene exp
SNPs
Whole Exome Seq
RNA seq

Generation of Hypothesies by PIs



"Drag & Drop"

High throughput analysis by Biostatistician & Bioinformatician

BCH Private network

integration
analysis

www.avillach-lab.hms.harvard.edu

Students / Postdocs


Cartik


Li


Romain


Qiu-Yue


Ombeline


Maxime


Antoine


Laurie


Laura


Haishuai


Mahdi


Niloofar


Alba


Joany


Carlos


Romina

Staff / Software developers


Jason


Ranjay


Gabor


Thomas


Jaspreet


Alex


Andrew


Andre


Sean


Anoush


Cassandra


Libby


Sophia


Alyssa

Alumni


Emmanuelle


Yuri


Samuel


Claire


Antoine


Michael


Sushma


Pei


Ephi


Jeremy

NIH Big Data to Knowledge (BD2K)

PIC-SURE

Open Source Research Infrastructure

amazon web services™

HIPAA
Health Insurance Portability and Accountability Act

Auth0

splunk>

docker
DOCKER DATACENTER

Windows Azure

IBM Cloud

SciDB

Google Cloud Platform

docker

i2b2

tranSMART

PIC-SURE
RESTful API

jupyter
Notebook

R

python

IJ

#!/bin/bash

cTAKES

gNOME

HARVARD MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics

NIH Big Data to Knowledge (BD2K)

PIC-SURE

amazon web services™

Harvard IRB security levels:

Level 5 - Extremely sensitive information

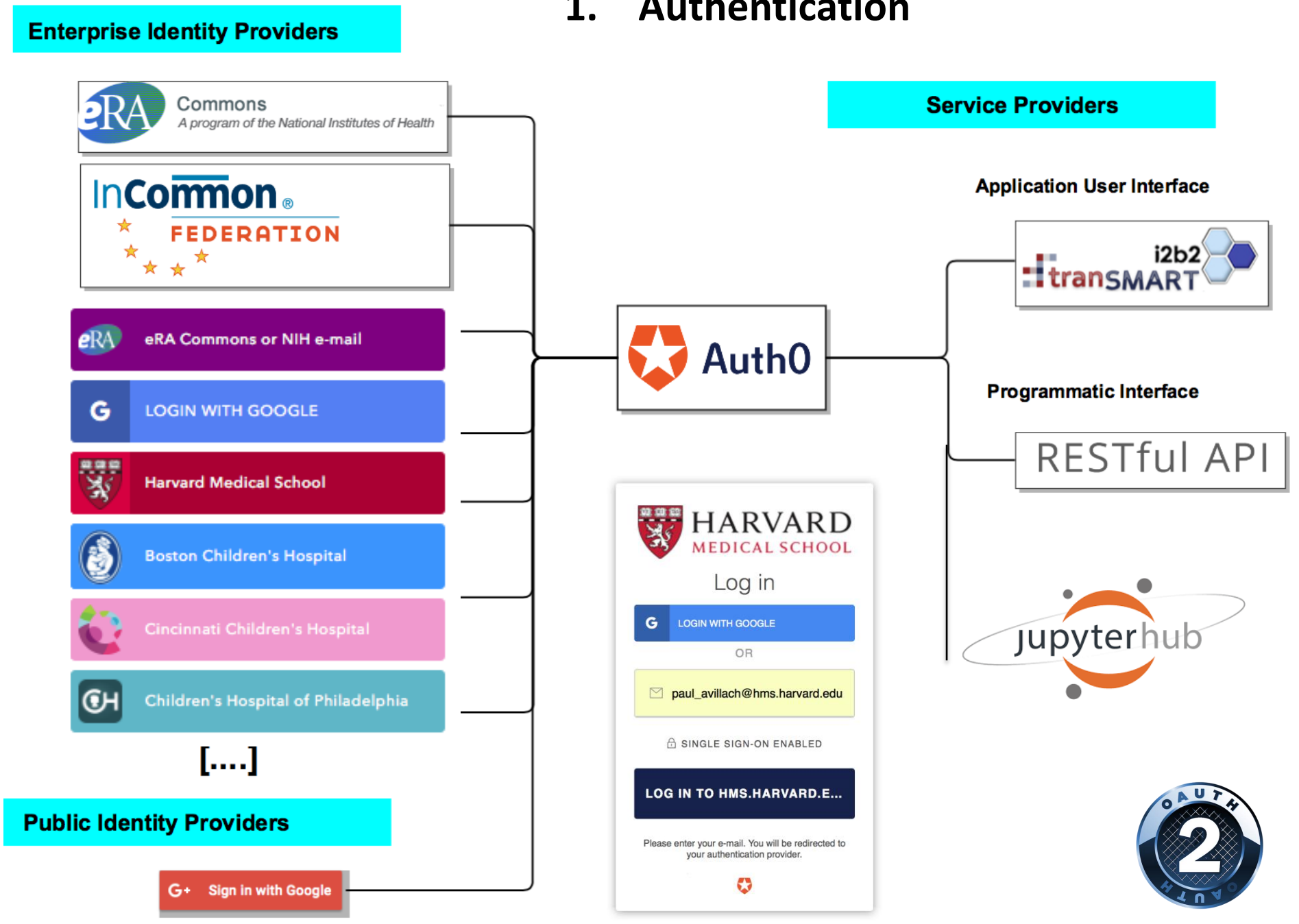
Level 4 – Very sensitive information

Level 3 – Sensitive, or Confidential information

Level 2 - Benign information to be held confidentially

Level 1 - Non-confidential research information

1. Authentication



Secured access control

1. **Authentication**

   Who are you?

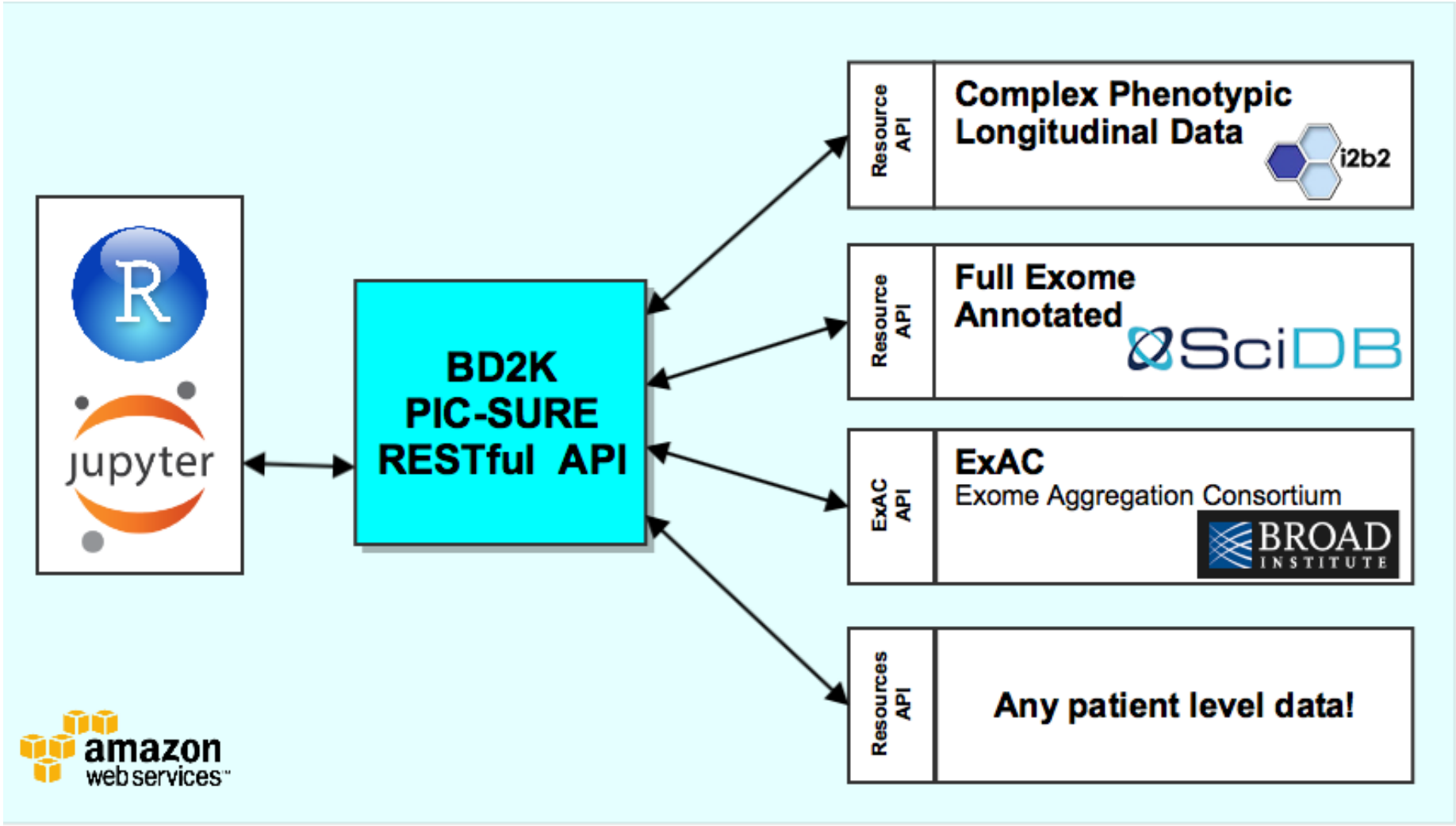
2. **Authorization**

   What are you allowed to do?

   Level 0 *: Authenticated BUT no access to data*

   Level 1 *: Aggregated data*

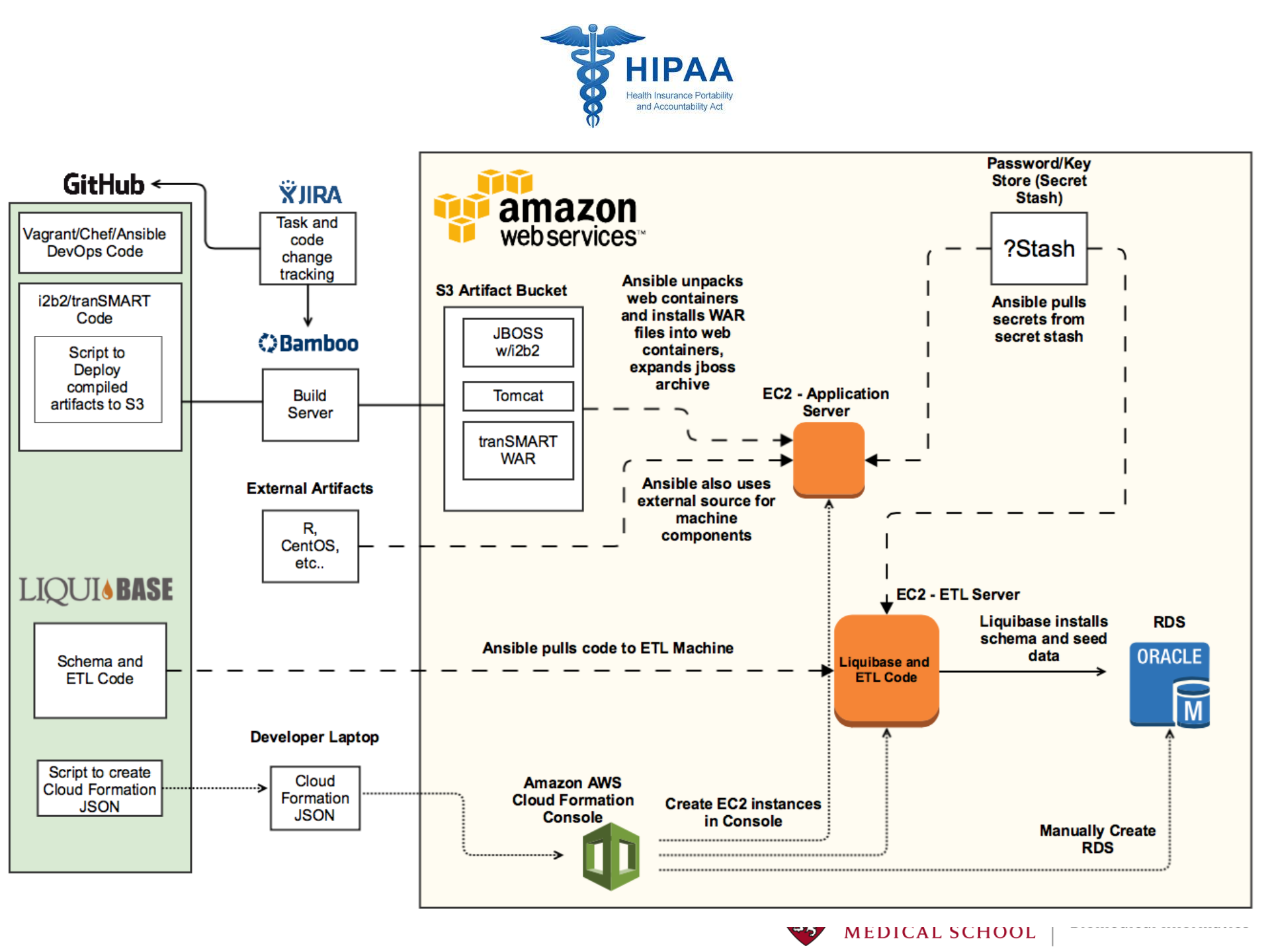
   Level 2 ***:*** *patient level data*



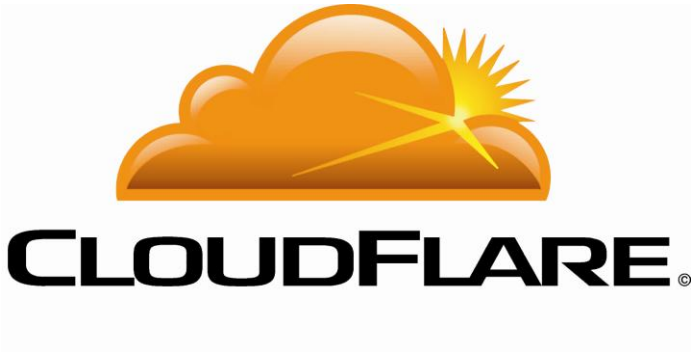
Play with PIC-SURE API: https://bd2k-picsure.hms.harvard.edu

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

2 Datathons


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

HIPPA Compliance on AWS



Amazon mystery solved: A typo took down a big chunk of the Internet

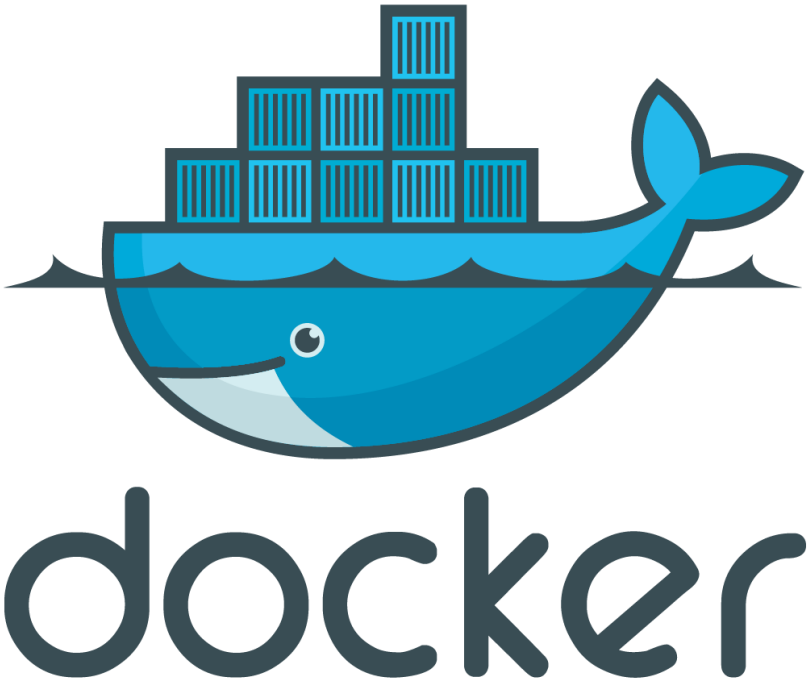
Elizabeth Weise , USATODAY Published 3:56 p.m. ET March 2, 2017 | **Updated 5:57 p.m. ET March 2, 2017**

(Photo: Amazon)

SAN FRANCISCO — The major outage that hit tens of thousands of websites using Amazon's AWS cloud computing service on Tuesday ends up having been the result of a simple typo — just one incorrectly-entered command.

The four-hour outage at Amazon Web Services' S3 system, a giant provider of backend services for close to 150,000 websites, caused disruptions, slowdowns and failure-to-load errors across the United States.


HIPAA
Health Insurance Portability
and Accountability Act


docker

NIH Commons Credits Portal

DOCKER DATACENTER

amazon web services™

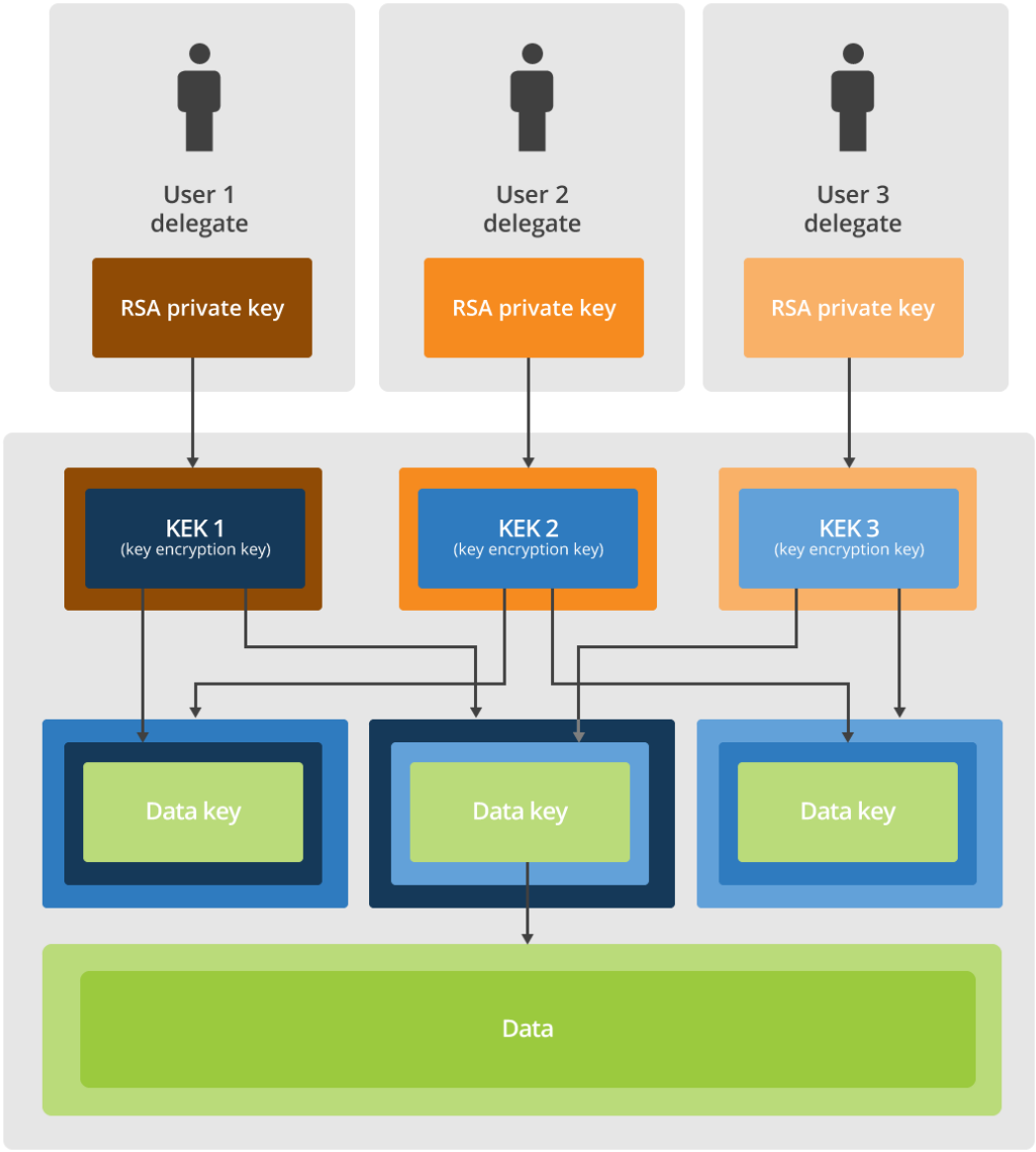
Google Cloud Platform

IBM Cloud

Windows Azure

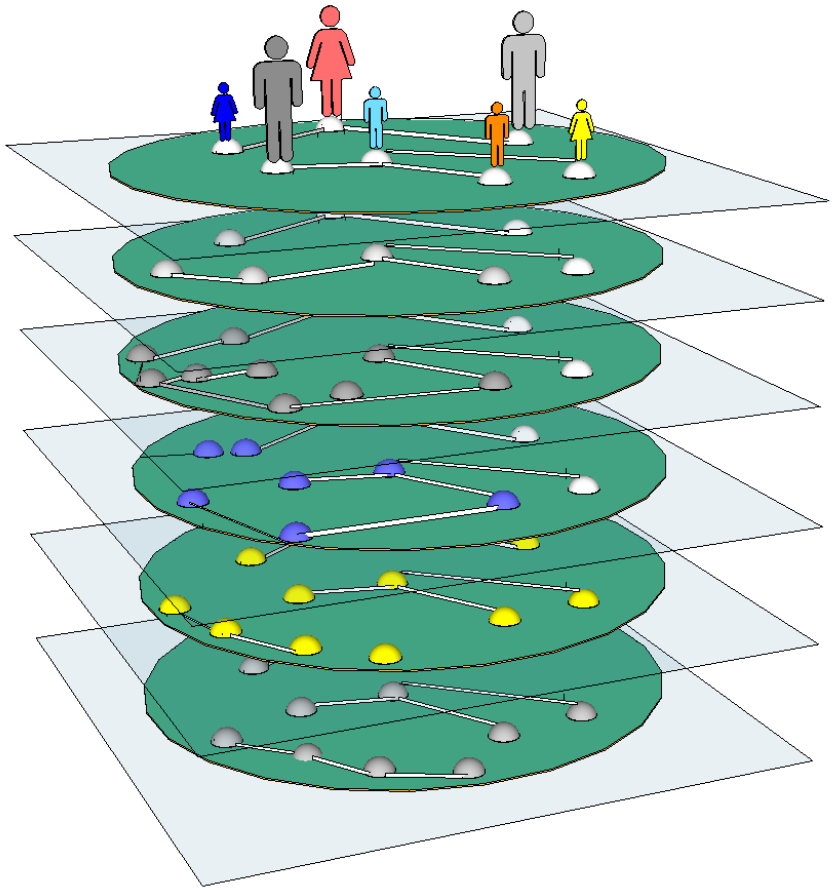
RedOctober

CLOUDFLARE®



UDN Patient centric information commons

i2b2/tranSMART/gNOME



ALL UDN data available

- Demographics
- Online registration Form
- PhenoTips
  - HPO clinical terms
  - Ethnicity
  - Diagnosis OMIM
- Knowledge extracted from referral letters (via NLP)
- Type of sequencing
- Candidate genes
- Candidate variants
- **Full Annotated VCF**
  - **524 Whole genomes**
  - **616 Whole exomes**
  - **343 probands**
  - **259 trios**
- + ….

NIH Big Data to Knowledge (BD2K)

PIC-SURE

amazon web services™

Modular Open Source Research platform

i2b2

tranSMART

PIC-SURE RESTful API

jupyter Notebook

R

python

IJ

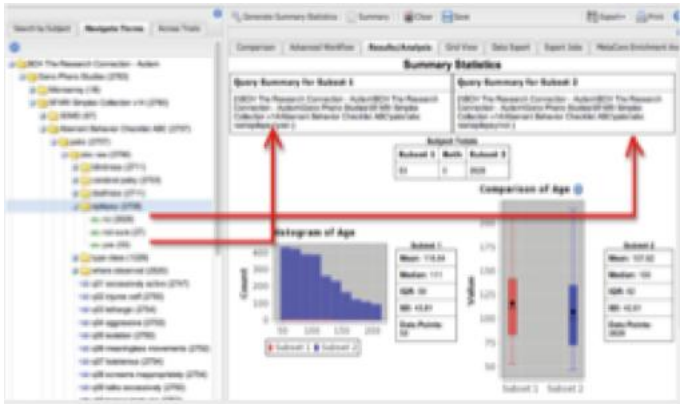
#!/bin/bash

cTAKES

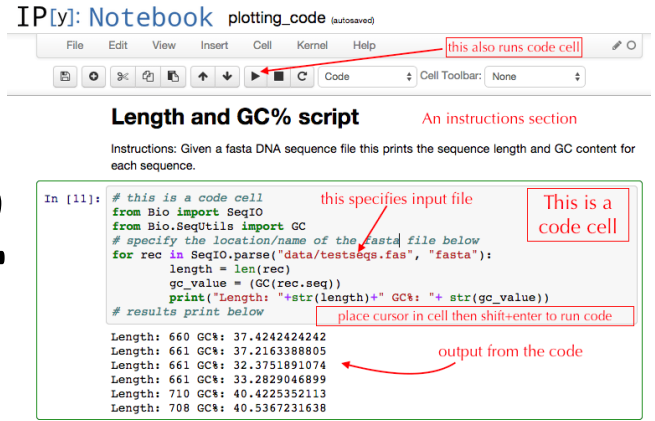
gNOME

SciDB

1



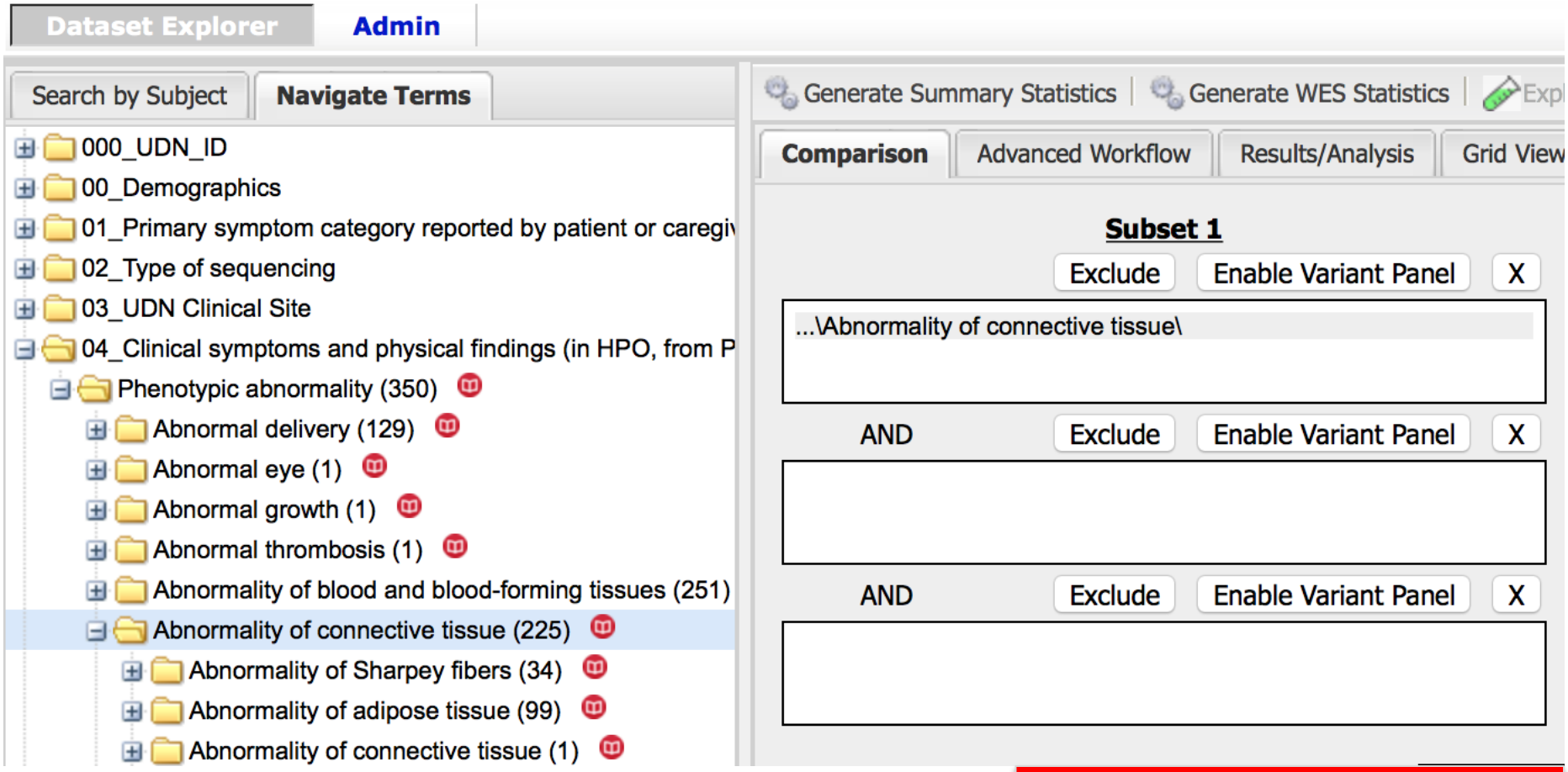
2

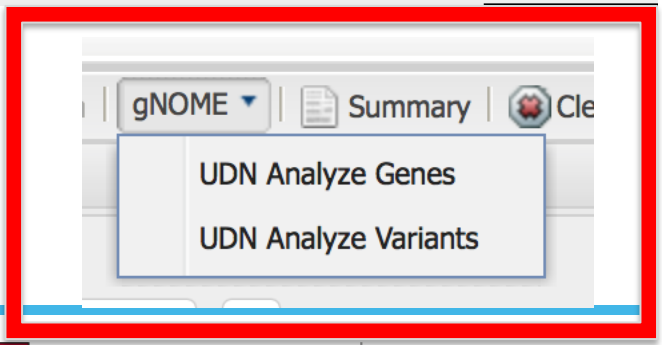


UDN338723

- 17 year old male with skeletal features suggestive of a **hereditary connective tissue disease**, also with proximal renal tubular acidosis, medullary nephrocalcinosis, hypophosphatemic rickets, polycythemia and chronic kidney disease, stage 3

- Marfan syndrome Confirmed
- FBN1 c.871G>T/p.E291X


Dataset Explorer
Admin
Search by Subject
Navigate Terms
000_UDN_ID
00_Demographics
01_Primary symptom category reported by patient or caregi
02_Type of sequencing
03_UDN Clinical Site
04_Clinical symptoms and physical findings (in HPO, from P
Phenotypic abnormality (350)
Abnormal delivery (129)
Abnormal eye (1)
Abnormal growth (1)
Abnormal thrombosis (1)
Abnormality of blood and blood-forming tissues (251)
Abnormality of connective tissue (225)
Abnormality of Sharpey fibers (34)
Abnormality of adipose tissue (99)
Abnormality of connective tissue (1)
Generate Summary Statistics
Generate WES Statistics
Comparison
Advanced Workflow
Results/Analysis
Grid View
Subset 1
Exclude
Enable Variant Panel
X
...\Abnormality of connective tissue\
AND
Exclude
Enable Variant Panel
X
AND
Exclude
Enable Variant Panel
X


gNOME
Summary
UDN Analyze Genes
UDN Analyze Variants

Analyze Variants

UDN_subset_de33b136c2a0c53 vs UDN_comple_subset_de33b136

Allele Frequency

Ancestry: All

Allele Frequency: ≤ 1% (rare)

Phenotype (powered by FindZebra)

Primary phenotype: connective AND

Secondary phenotype: AND

Tertiary phenotype:

Use gene ranked in top: 5 %

For each phenotype, please enter as many terms (separated by comma) describing the phenotype as possible. A single word (e.g., headache) or multipel words separated by space (e.g., neck pain) can be a single term. For each phenotype, only genes highly ranked by relevancy to the phenotype will be considered. Only the genes associated with all of the above phenotypes will be shown.

Functional Impact

Gene model: RefSeq

Gene impact: nonsynonymous

SIFT: any

PolyPhen2: any

Condel: any

Statistical Test Parameters

(not applied to per genome enrichment)

p-value threshold ≤ 1

$N_{\text{group A genomes}} \geq$ 1

$N_{\text{group B genomes}} \leq$ 50

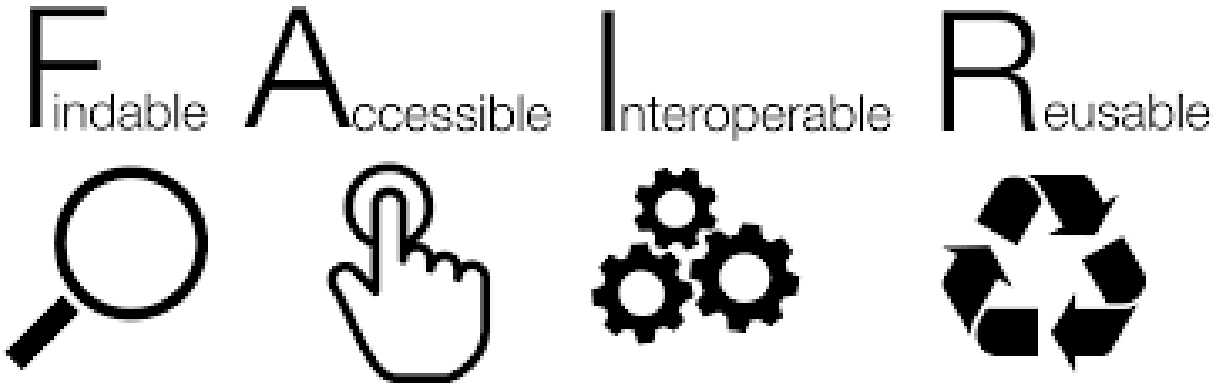
271 candidate variants (in 86 transcripts and 76 genes).

Transcript ID	Gene symbol (ID)	CDS size (bps)	Chromosome	Variant start	Variant end	Reference sequence		Wilcoxon p-value	number of case genomes	number of control genomes
NM_001127707	SERPINA1	1,257	chr14	94,845,944	94,845,944	C	T	0	34	46
NM_001127707	SERPINA1	1,257	chr14	94,845,914	94,845,914	T	A	0	32	40
NM_001127707	SERPINA1	1,257	chr14	94,845,886	94,845,886	-	G	0	23	30
NM_001127707	SERPINA1	1,257	chr14	94,844,968	94,844,968	T	C	0	20	30
NM_001127707	SERPINA1	1,257	chr14	94,845,885	94,845,885	-	TA	0.005	13	23
NM_032470	TNXB	2,022	chr6	32,010,523	32,010,523	A	G	0.262	12	40
NM_032470	TNXB	2,022	chr6	32,010,732	32,010,732	T	G	0.043	12	28
NM_000492	CFTR	4,443	chr7	117,188,877	117,188,877	G	T	0.004	11	17
NM_032470	TNXB	2,022	chr6	32,010,572	32,010,572	G	T	0.657	10	46
NM_000257	MYH7	5,808	chr14	23,889,445	23,889,445	-	G	0.779	9	47

[....]

NM_024312	GNPTAB	3,771	chr12	102,190,521	102,190,521	C	T	0.279	1	2
NM_000257	MYH7	5,808	chr14	23,884,281	23,884,281	C	T	0.023	1	0
NM_000257	MYH7	5,808	chr14	23,887,578	23,887,578	C	T	0.023	1	0
NM_000257	MYH7	5,808	chr14	23,892,910	23,892,910	A	G	0.023	1	0
NM_000257	MYH7	5,808	chr14	23,894,554	23,894,554	C	T	0.023	1	0
NM_001127707	SERPINA1	1,257	chr14	94,847,386	94,847,386	G	A	0.399	1	3
NM_000138	FBN1	8,616	chr15	48,704,813	48,704,813	C	T	0.023	1	0
NM_000138	FBN1	8,616	chr15	48,714,160	48,714,160	G	A	0.023	1	0
NM_000138	FBN1	8,616	chr15	48,782,072	48,782,072	T	C	0.143	1	1
NM_001009944	PKD1	12,912	chr16	2,140,294	2,140,294	C	T	0.497	1	4
NM_001009944	PKD1	12,912	chr16	2,153,345	2,153,345	C	T	0.399	1	3
NM_001009944	PKD1	12,912	chr16	2,157,984	2,157,984	G	A	0.143	1	1
NM_001009944	PKD1	12,912	chr16	2,158,419	2,158,419	G	A	0.279	1	2
NM_001009944	PKD1	12,912	chr16	2,158,869	2,158,869	G	A	0.023	1	0
NM_001009944	PKD1	12,912	chr16	2,159,557	2,159,557	C	T	0.497	1	4

OPEN DATA



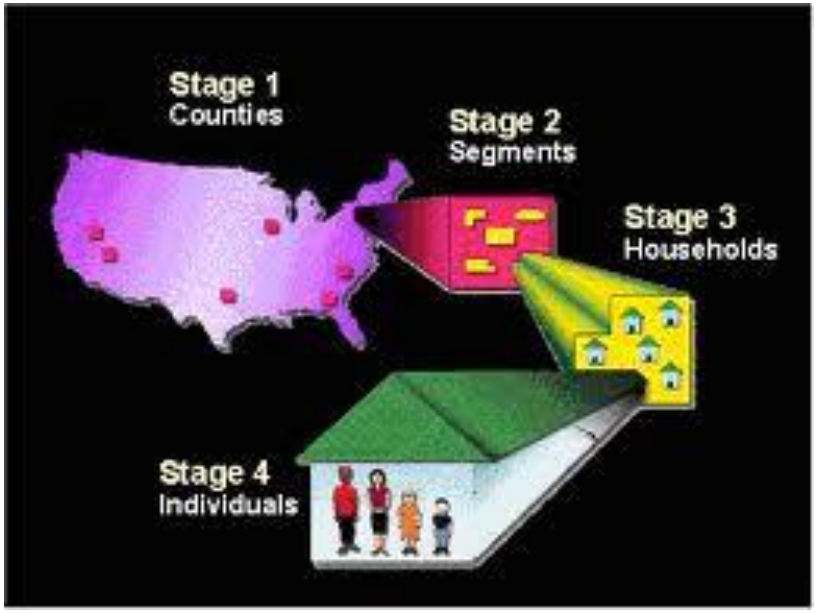
CDC National Health and Nutrition Examination Survey

https://nhanes.hms.harvard.edu

Gold standard resource for *phenotypes and exposure* information: National Health and Nutrition Examination Survey

since the 1960s
now biannual: 1999 onwards
5000 participants per year

1999-2006:
N=41,474
1,181 total measurements!



>250 exposures (serum + urine)
>85 quantitative clinical traits
(e.g., serum glucose, lipids, body mass index)
nutritional and health status interviews

1 http://www.cdc.gov/nchs/nhanes.htm

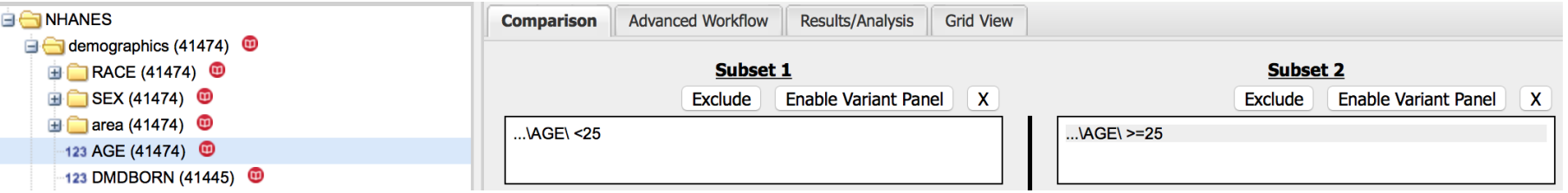
CDC National Health and Nutrition Examination Survey

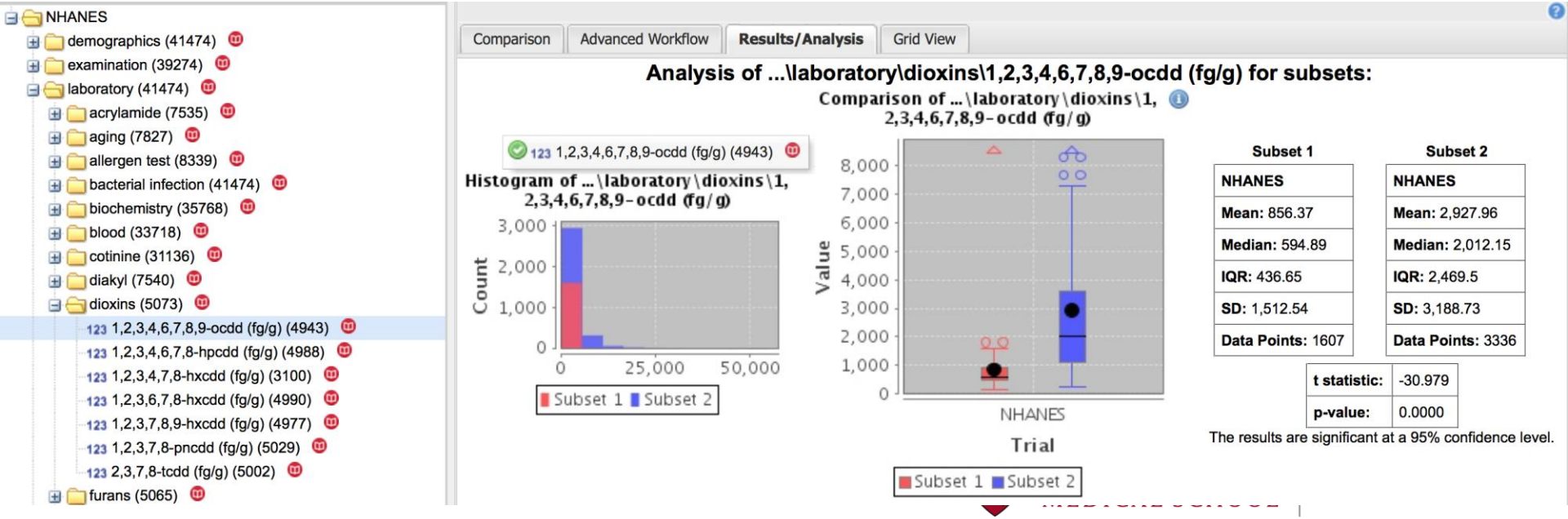
41k patients

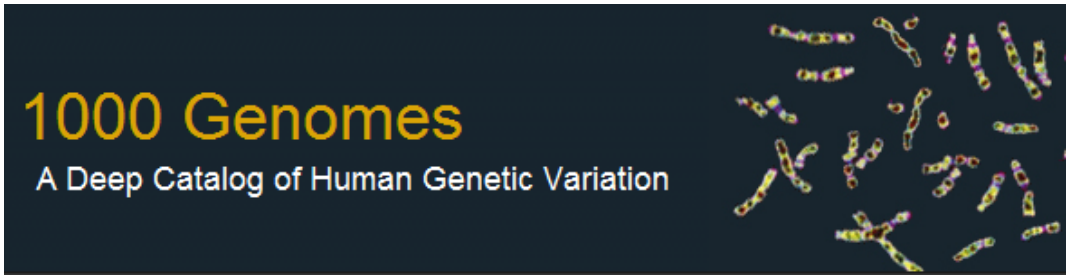
2k patient level environmental variables

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

10 min video https://vimeo.com/182576739



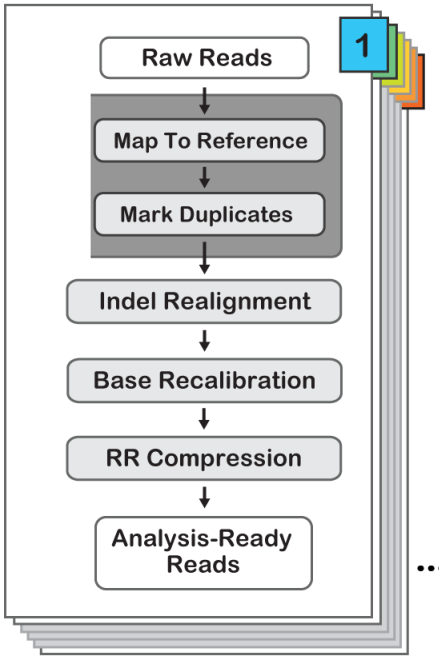


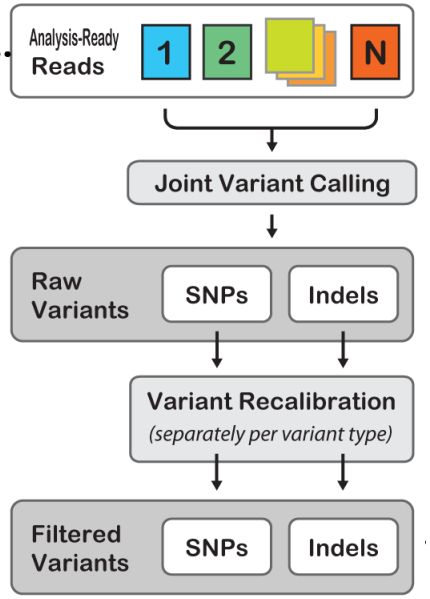


- 56 full exomes with Phenotypic data from 1000 Genome project:
- *no registration at all*

https://demo-ngs.hms.harvard.edu

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

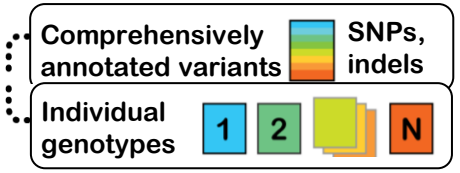


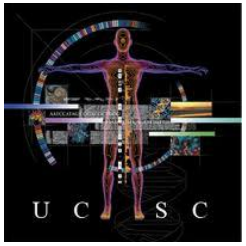


gatk

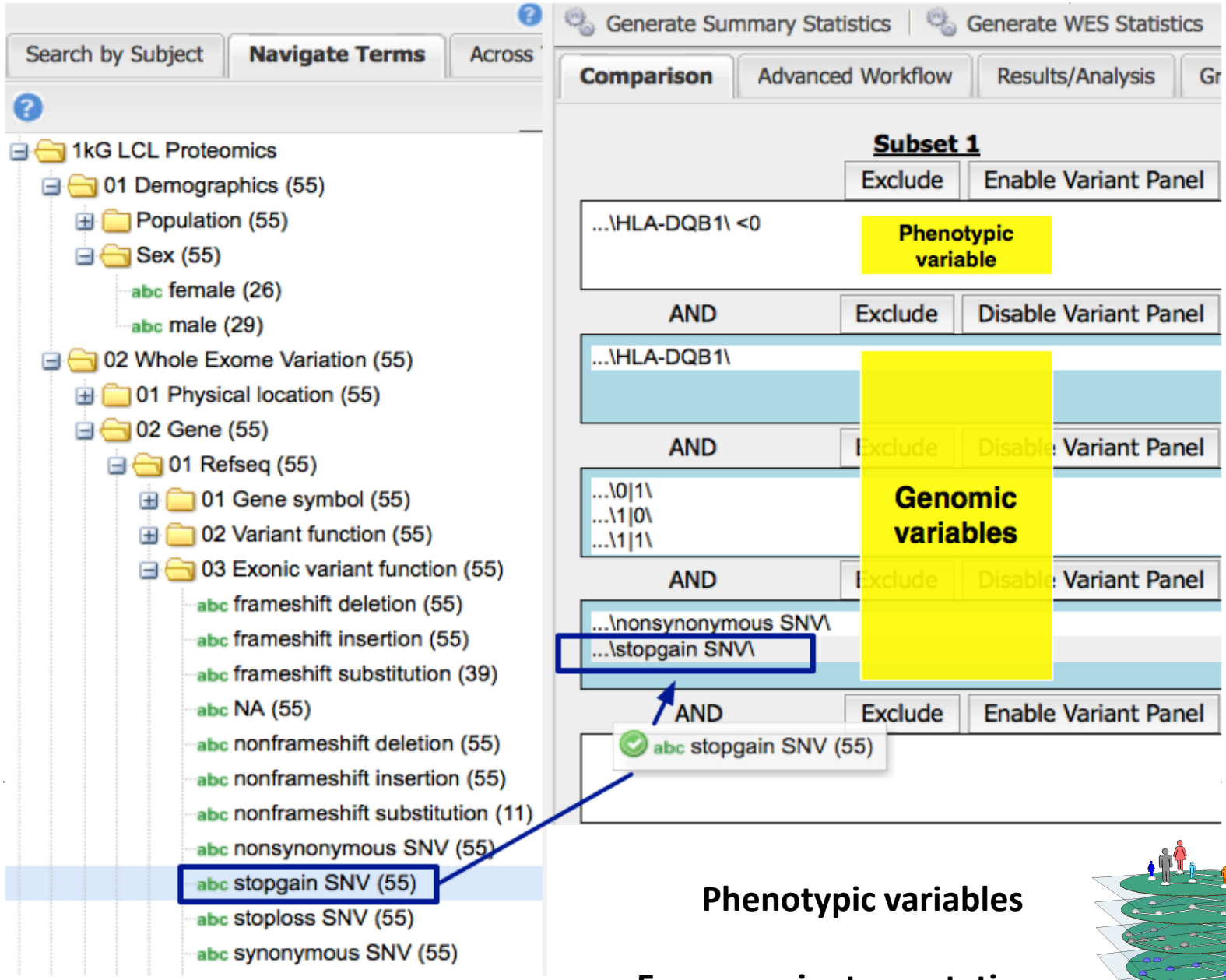
https://demo-ngs.hms.harvard.edu

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



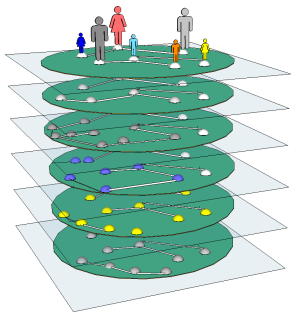


ANNOVAR



Phenotypic variables

Exome variant annotations



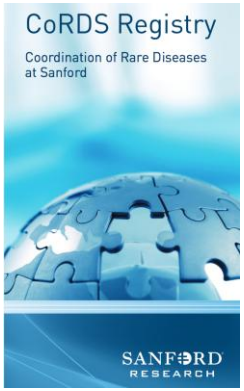
https://demo-ngs.hms.harvard.edu

NIH 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)



National Ataxia Foundation

INTRACRANIAL HYPERTENSION
RESEARCH FOUNDATION
CONTACT

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

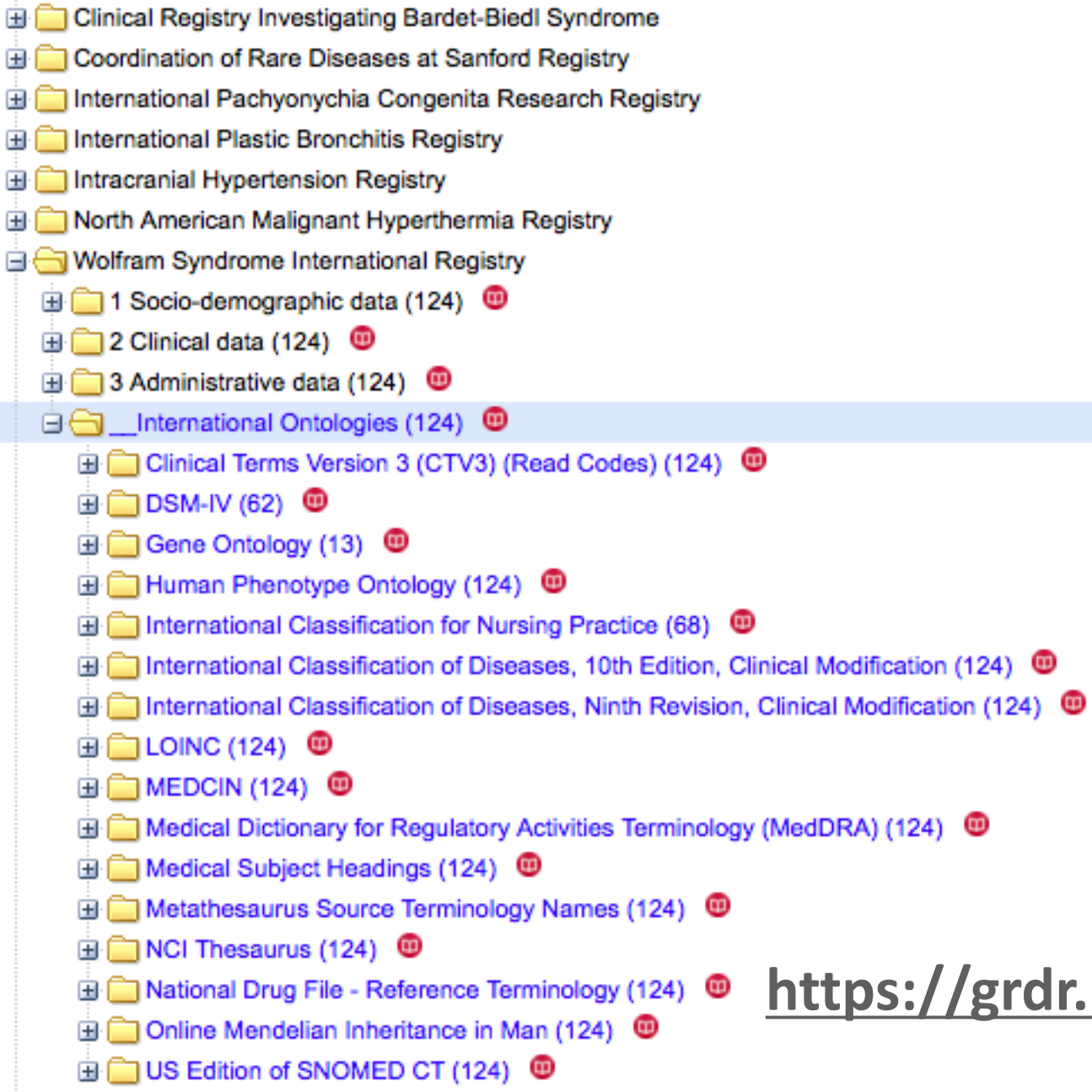
CdLS Foundation
Cornelia de Lange Syndrome Foundation, Inc.

https://grdr.hms.harvard.edu

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

Rare disease registry	Patient count	Variables	Version
Clinical Registry Investigating Bardet-Biedl Syndrome	180	708	**V4**
International Pachyonychia Congenita Research Registry	569	496	**V3**
International Plastic Bronchitis registry	66	63	V2
Intracranial Hypertension Registry	963	77	V4
North American Malignant Hyperthermia Registry	2,107	162	**V4**
Wolfram Syndrome International Registry	124	580	V2
Coordination of Rare Diseases at Sanford Registry	1,218	46	V2
including:	*including:*	*Additional data:*	
National Ataxia Foundation	*487*	*16*	*V1*
International WAGR syndrome association	*50*	*380*	*V1*
Cornelia De Lange Syndrome Registry	*61*	*486*	*V1*
Total	***5,227***	***3,014***	
Rare Cancer Registry (soon)	800	1,800	
Centronuclear and myotubular myopathy (soon)			

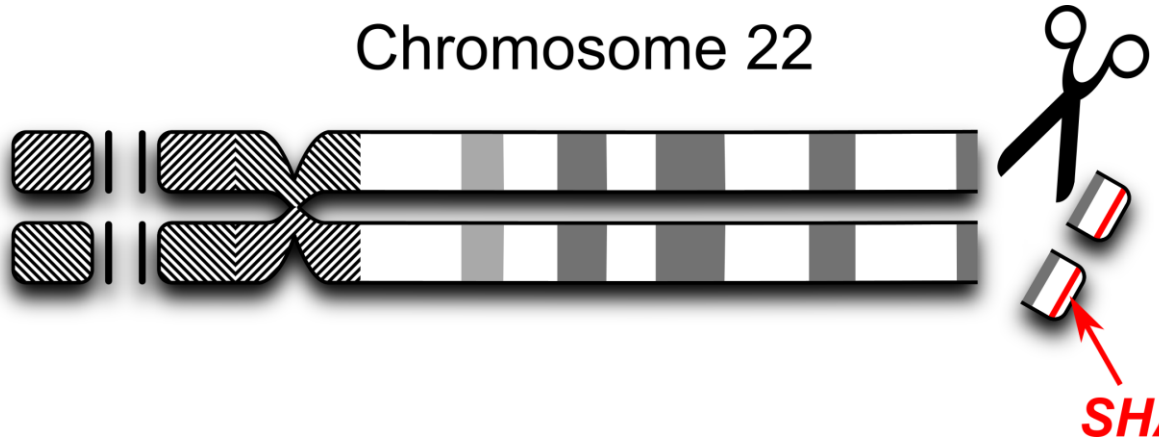
https://grdr.hms.harvard.edu



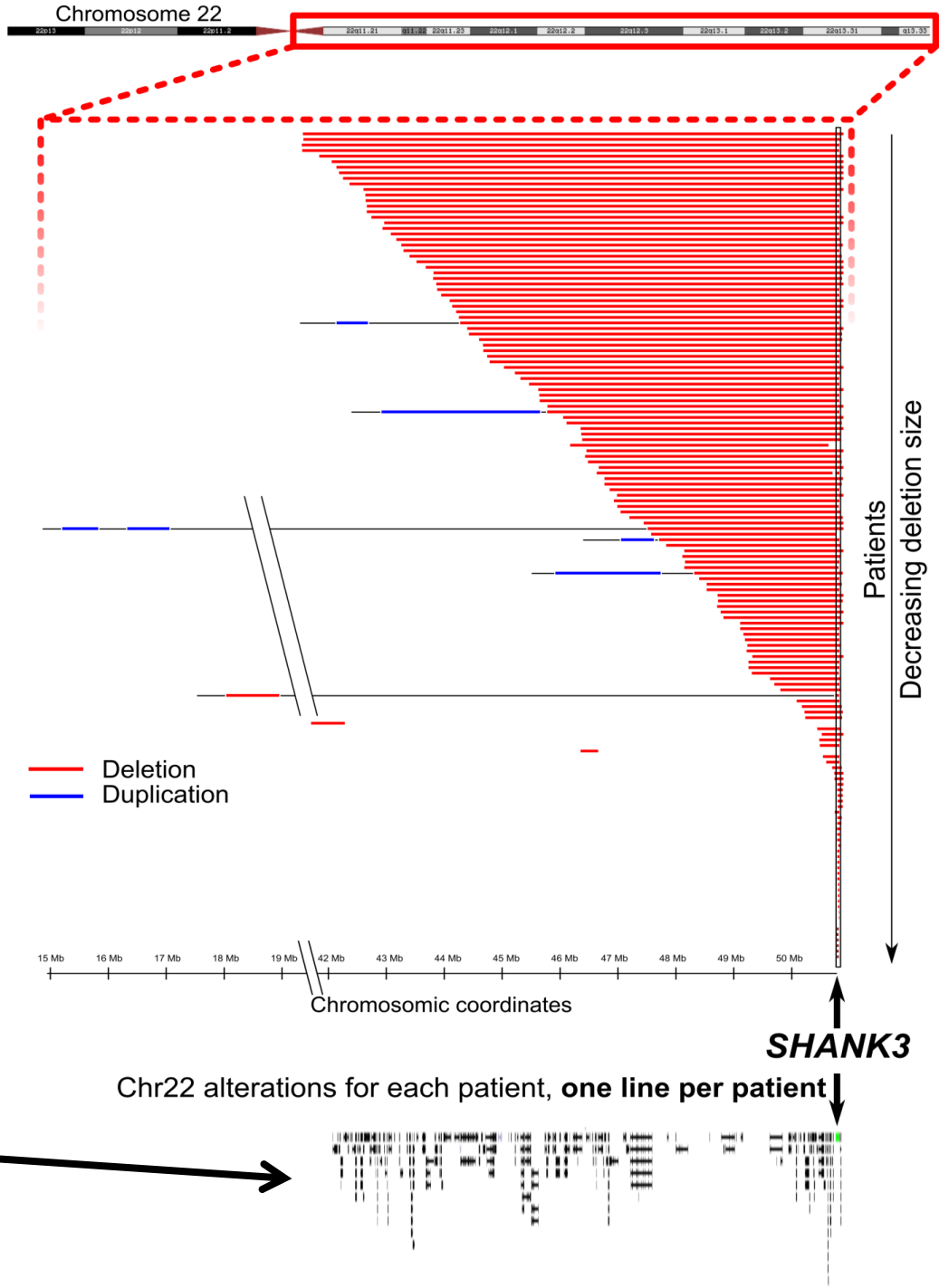
https://grdr.hms.harvard.edu

Phelan McDermid Syndrome

- Extremely rare genetic disease: **~1100 diagnosed patients worldwide today**
- Autistic traits, **intellectual deficiency**, slight dysmorphic features
- Also called **deletion 22q13 syndrome**
- Caused by deletions of the terminus of chromosome 22



Heterogeneity of the genetic material alterations



1 to 140 affected genes

Heterogeneity of the phenotypes

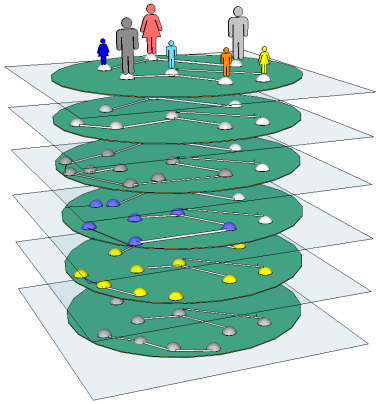
Sources :
globalgenes.org
autismspeaks.org
sfari.org

All organs can be affected:

- Neuro-developmental
- Facial dysmorphic features
- GERD (Gastro-Esophagal Reflux)
- Renal problems
- Lax joints
- Dysplastic toenails
- Congenital cardiac diseases
- …

Deep phenotyping

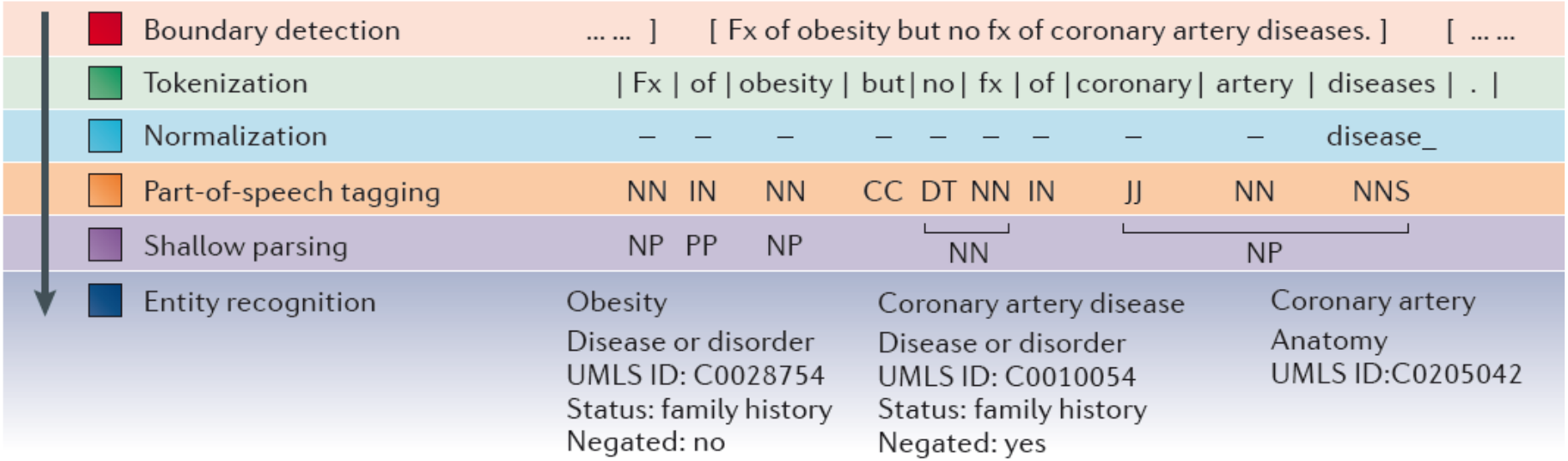
Knowledge from Clinical Notes



cTAKES™

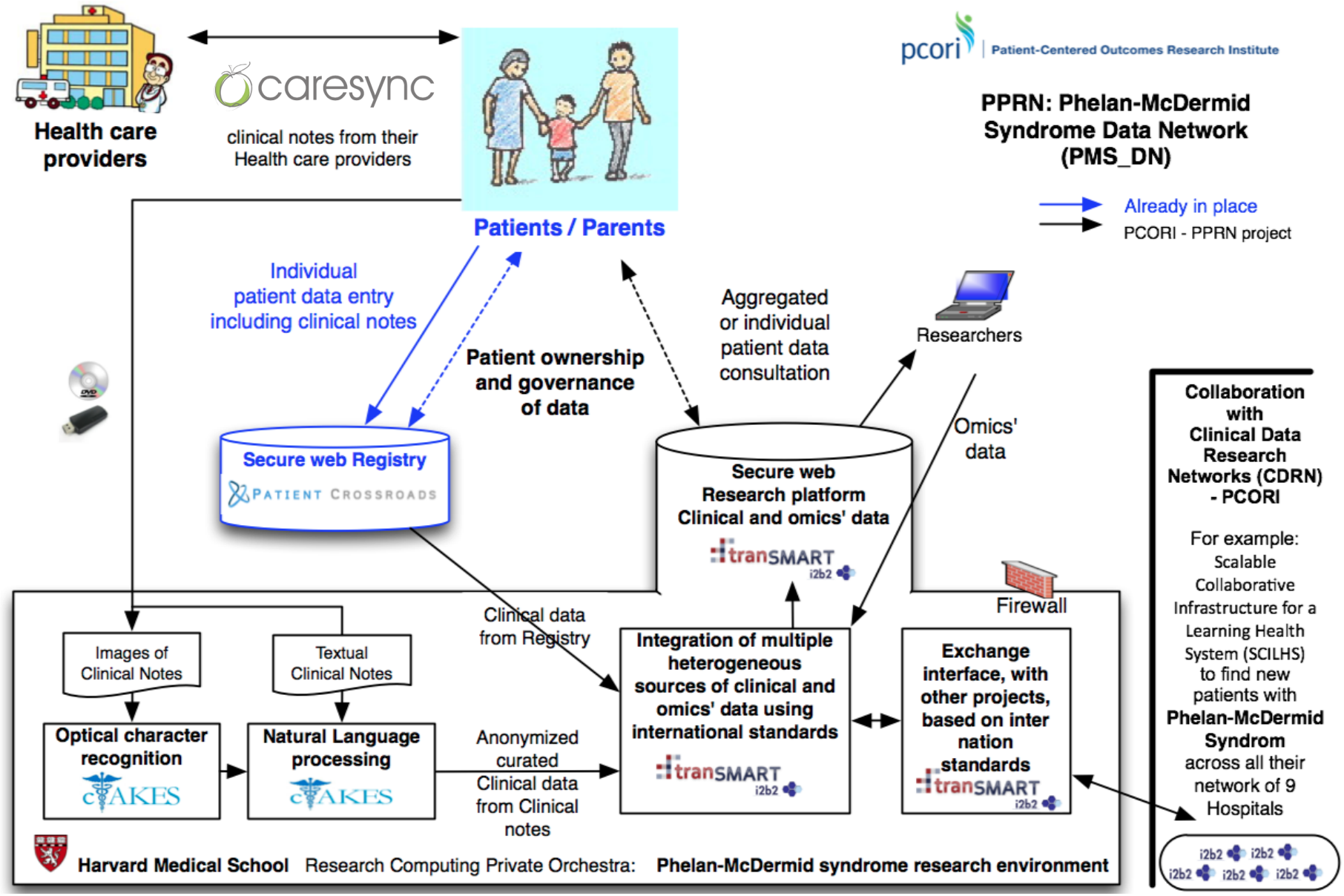
The Apache Software Foundation
http://www.apache.org/

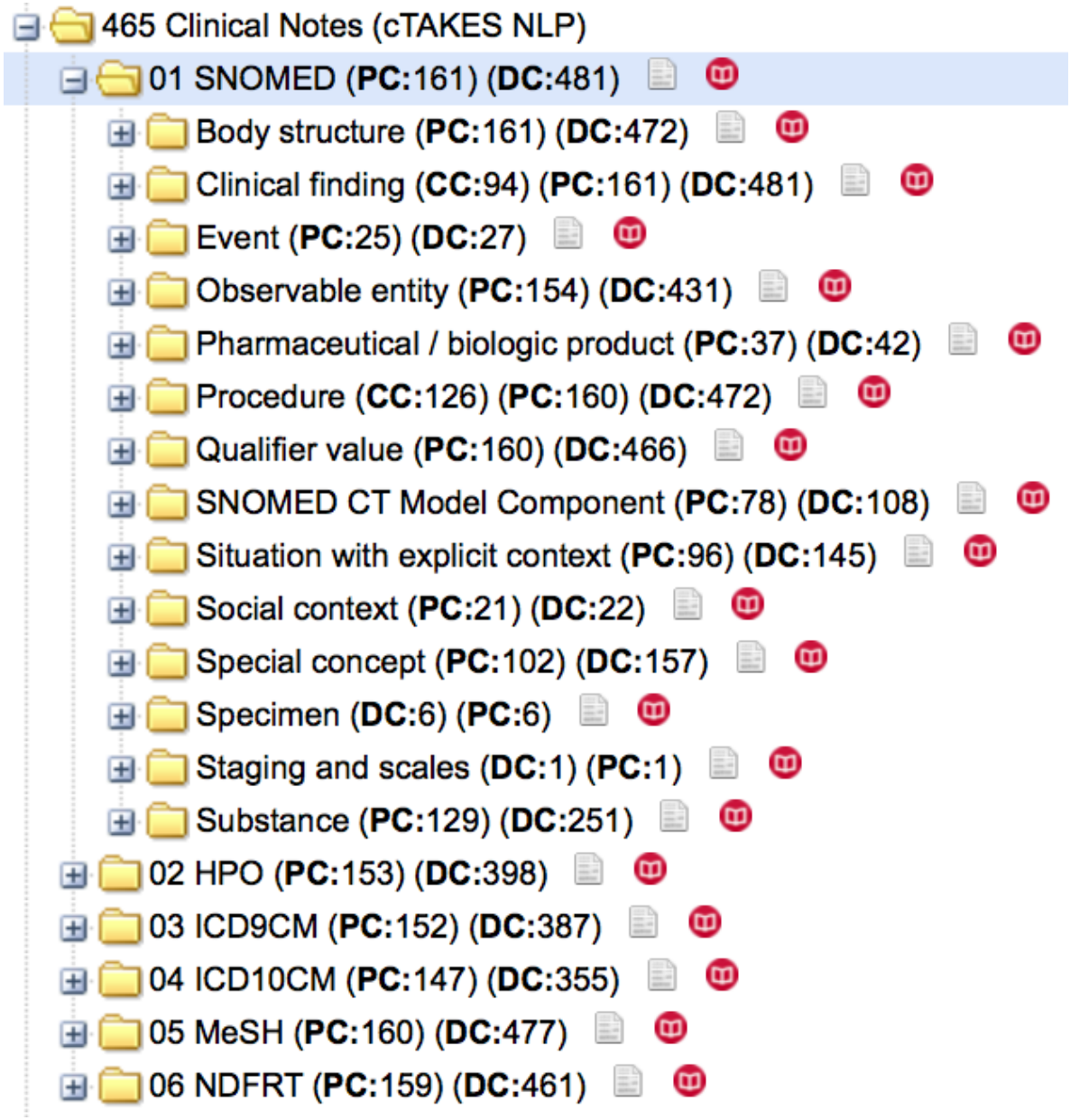
Box 1 | **Natural language processing**



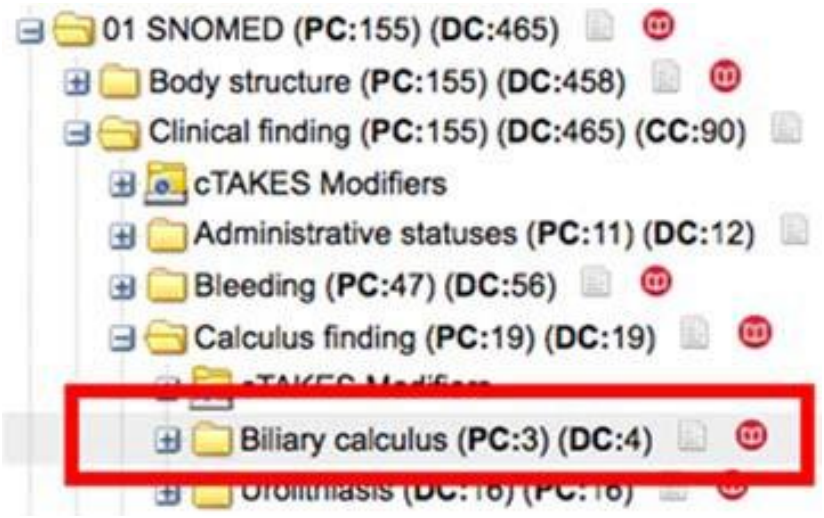
Peter B. Jensen, Lars J. Jensen and Søren Brunak, Nat Rev Genet. 2012

caresync

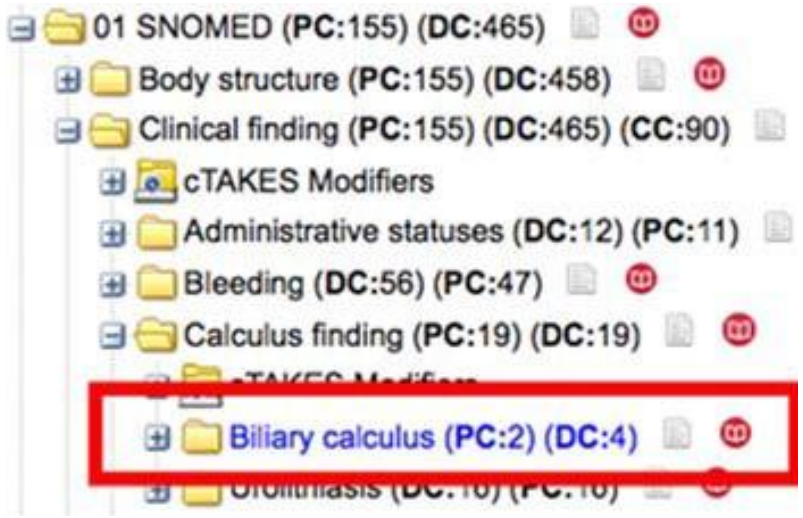

Health care providers
caresync
clinical notes from their Health care providers
Patients / Parents
pcori
Patient-Centered Outcomes Research Institute
PPRN: Phelan-McDermid Syndrome Data Network (PMS_DN)
Already in place
PCORI - PPRN project
Individual patient data entry including clinical notes
Patient ownership and governance of data
Aggregated or individual patient data consultation
Researchers
Omics' data
Secure web Registry
PATIENT CROSSROADS
Secure web Research platform Clinical and omics' data
tranSMART
i2b2
Firewall
Clinical data from Registry
Images of Clinical Notes
Textual Clinical Notes
Optical character recognition
cTAKES
Natural Language processing
cTAKES
Anonymized curated Clinical data from Clinical notes
Integration of multiple heterogeneous sources of clinical and omics' data using international standards
tranSMART
i2b2
Exchange interface, with other projects, based on inter nation standards
tranSMART
i2b2
Harvard Medical School Research Computing Private Orchestra: Phelan-McDermid syndrome research environment
Collaboration with Clinical Data Research Networks (CDRN) - PCORI
For example: Scalable Collaborative Infrastructure for a Learning Health System (SCILHS) to find new patients with Phelan-McDermid Syndrom across all their network of 9 Hospitals
i2b2 i2b2 i2b2 i2b2 i2b2


465 Clinical Notes (cTAKES NLP)
01 SNOMED (PC:161) (DC:481)
Body structure (PC:161) (DC:472)
Clinical finding (CC:94) (PC:161) (DC:481)
Event (PC:25) (DC:27)
Observable entity (PC:154) (DC:431)
Pharmaceutical / biologic product (PC:37) (DC:42)
Procedure (CC:126) (PC:160) (DC:472)
Qualifier value (PC:160) (DC:466)
SNOMED CT Model Component (PC:78) (DC:108)
Situation with explicit context (PC:96) (DC:145)
Social context (PC:21) (DC:22)
Special concept (PC:102) (DC:157)
Specimen (DC:6) (PC:6)
Staging and scales (DC:1) (PC:1)
Substance (PC:129) (DC:251)
02 HPO (PC:153) (DC:398)
03 ICD9CM (PC:152) (DC:387)
04 ICD10CM (PC:147) (DC:355)
05 MeSH (PC:160) (DC:477)
06 NDFRT (PC:159) (DC:461)

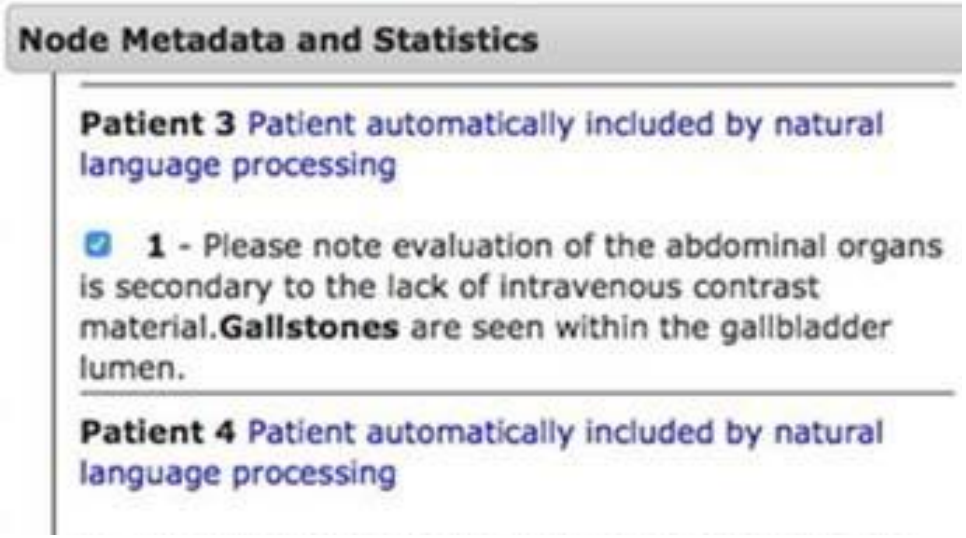
BEFORE validation



AFTER validation



Pop-up validation window



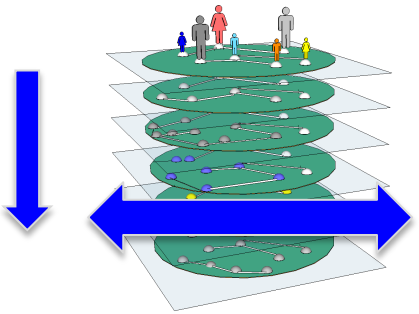
Objective:

Identify phenotypes linked

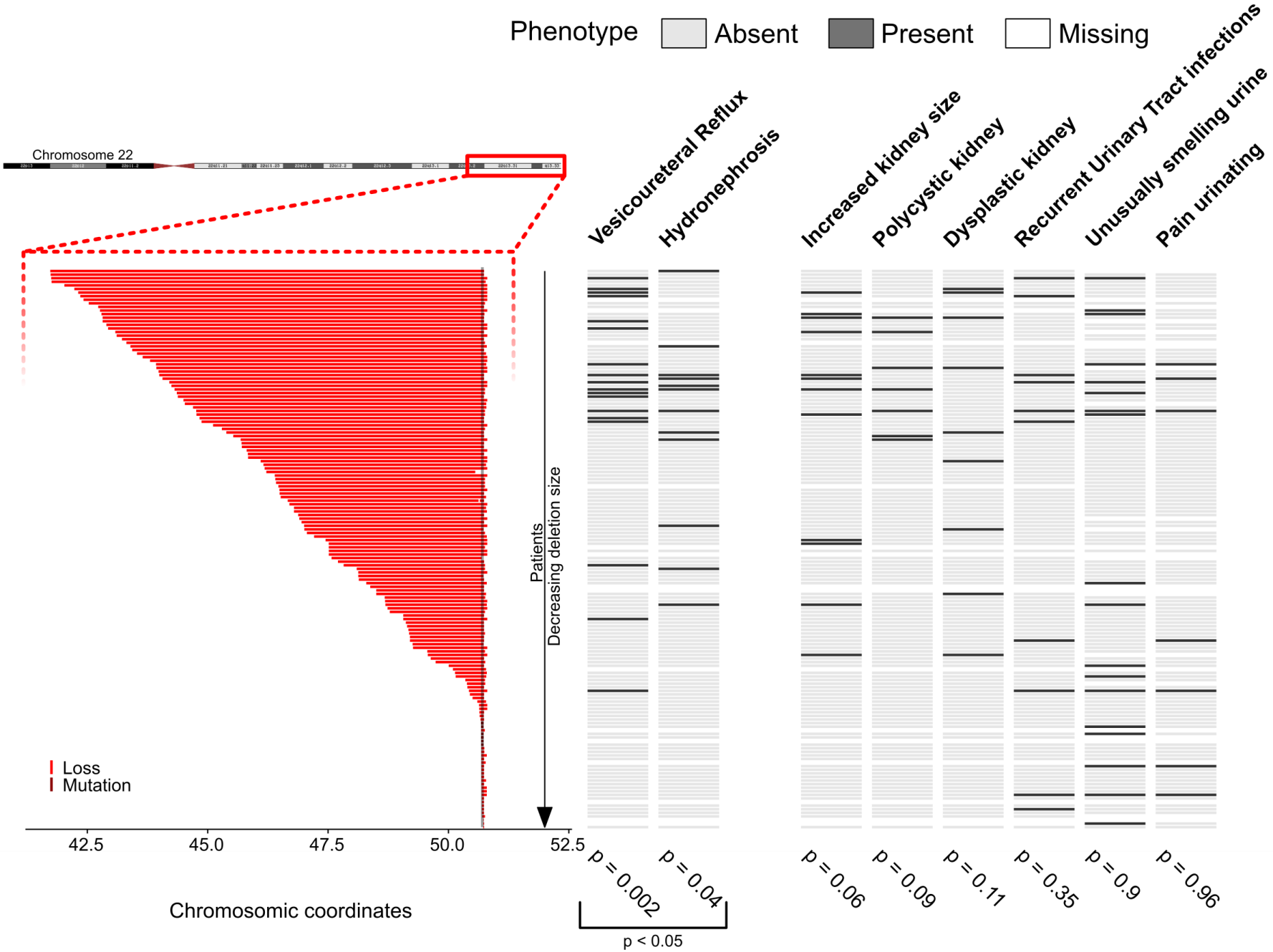
with the **deletion <u>of other genes</u>** than *SHANK3*

Chr deletion size

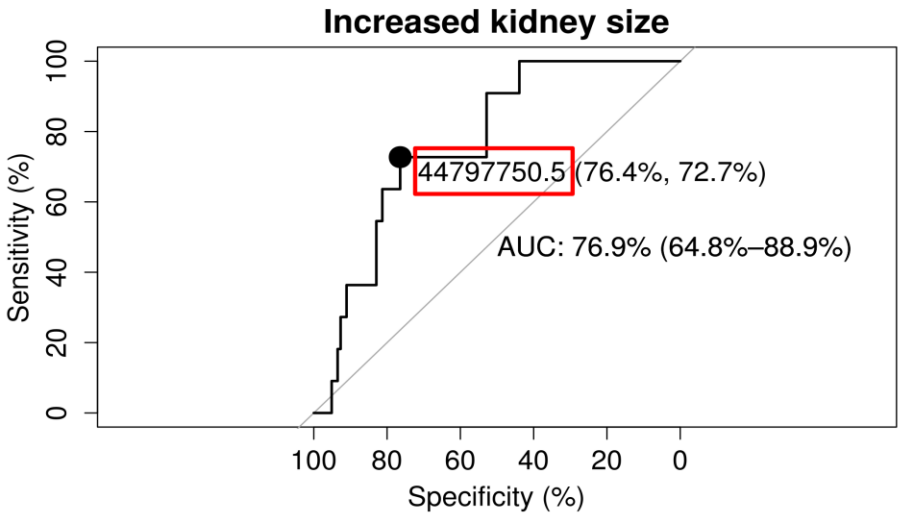
Phenotypes

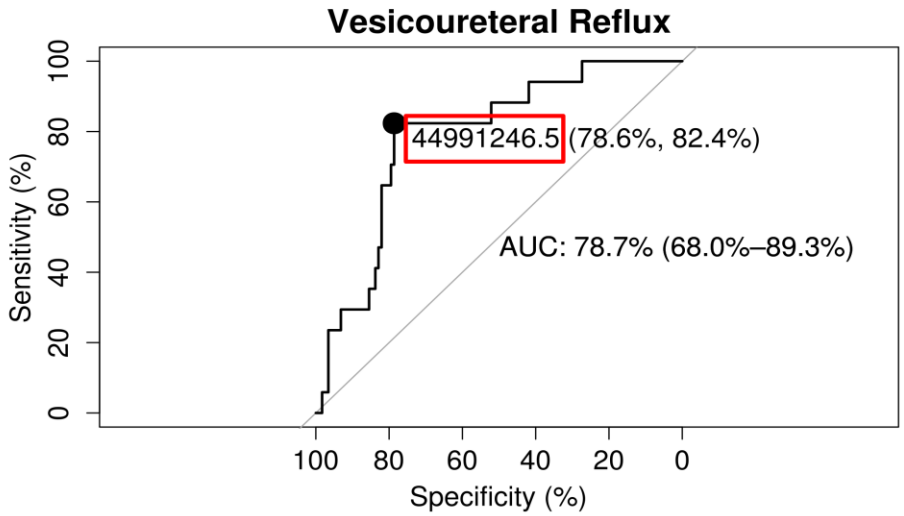


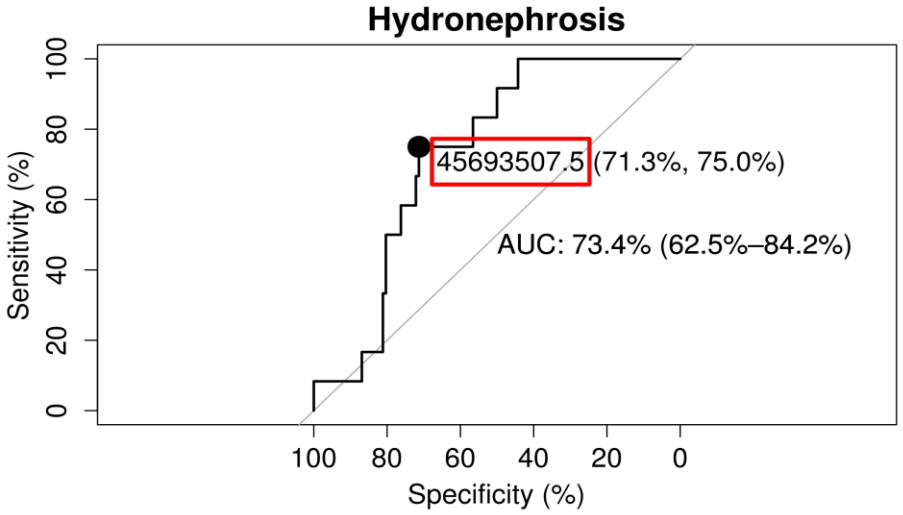
Results – Kidney malformations associated **with other genes** that *SHANK3*

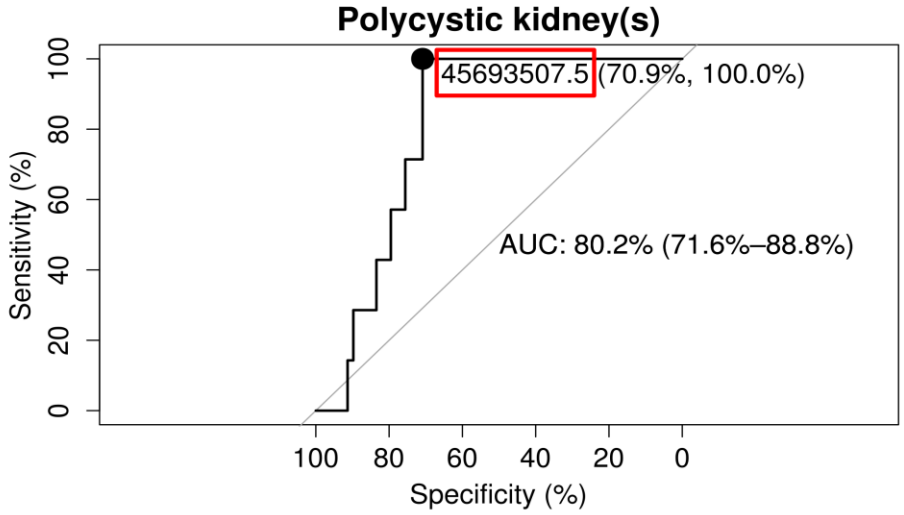


Panel A

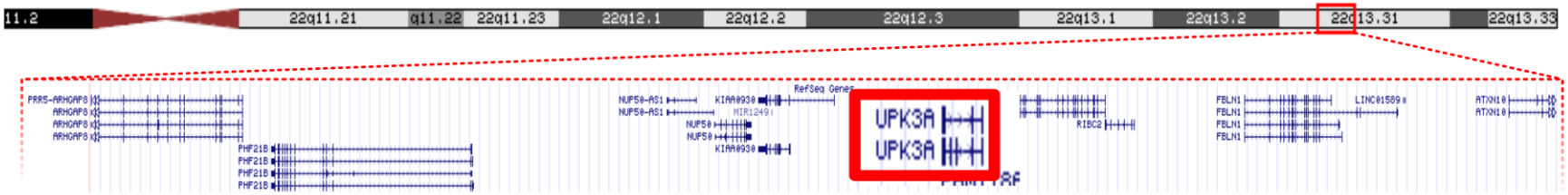




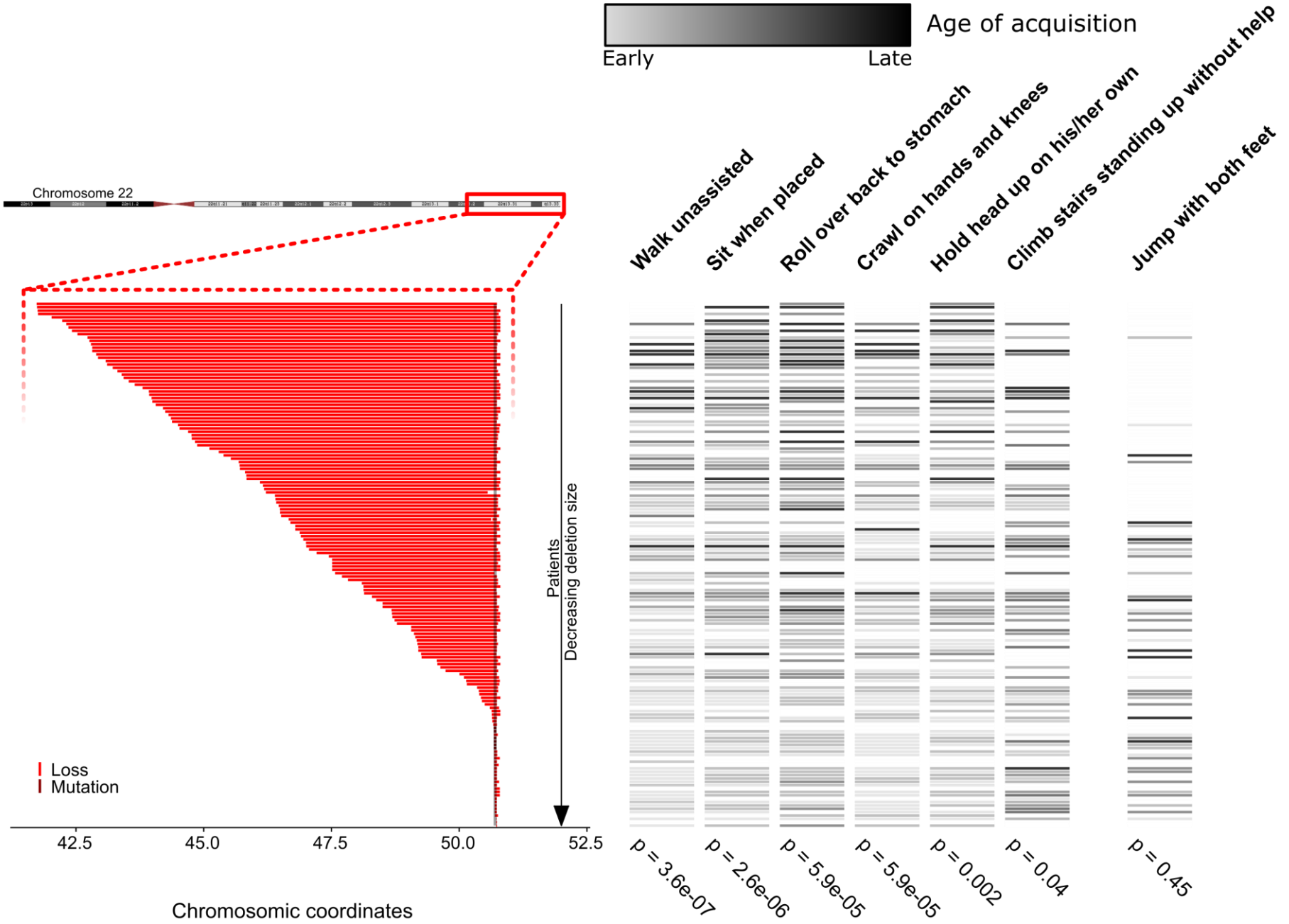




Panel B



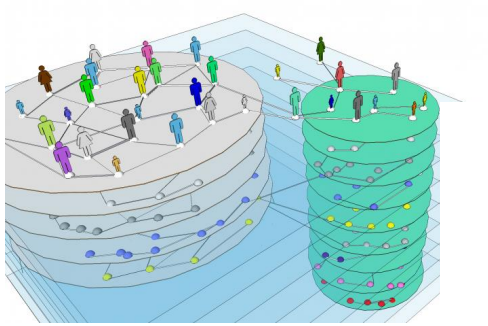
Results – Gradient of gross motor delays showing a cumulative effect



i2b2 1.7.x

1.0_HOTFIX

tranSMART FOUNDATION



i2b2 tranSMART

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

Harvard – DBMI

AvillachLab improvements

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

1.7.x

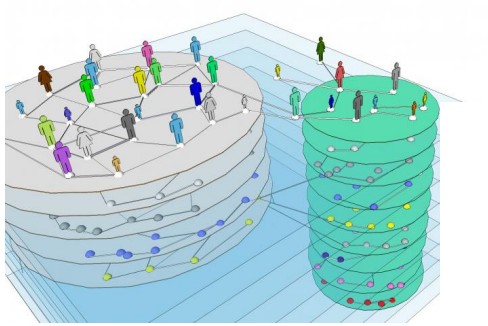
i2b2

docker

16.2 ~~(17.12)~~

tranSMART FOUNDATION

i2b2 tranSMART



open source

Harvard – DBMI

AvillachLab improvements

An i2b2/tranSMART project management committee

If interested to join, email us:

Diane:
dkeogh@i2b2foundation.org

Paul:
paul_avillach@hms.harvard.edu

- **Play with i2b2/tranSMART UI:**
- https://grdr.hms.harvard.edu
- https://nhanes.hms.harvard.edu
- https://demo-ngs.hms.harvard.edu
- https://pmsdn.hms.harvard.edu *(ASD researcher only)*

- **Play with BD2K PIC-SURE RESTful API**
- http://bd2k-picsure.hms.harvard.edu
- http://exac.hms.harvard.edu

www.avillach-lab.hms.harvard.edu

Students / Postdocs


Cartik


Li


Romain


Qiu-Yue


Ombeline


Maxime


Antoine


Laurie


Laura


Haishuai


Mahdi


Niloofar


Alba


Joany


Carlos


Romina

Staff / Software developers


Jason


Ranjay


Gabor


Thomas


Jaspreet


Alex


Andrew


Andre


Sean


Anoush


Cassandra


Libby


Sophia


Alyssa

Alumni


Emmanuelle


Yuri


Samuel


Claire


Antoine


Michael


Sushma


Pei


Ephi


Jeremy

Postdoc? / Mobilité?

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

www.avillach-lab.hms.harvard.edu

Paul_Avillach.hms.harvard.edu