Translating a Trillion Points of Data into Diagnostics, Therapies and New Insights in Health and Disease

Atul Butte, MD, PhD

atul.butte@ucsf.edu

@atulbutte

Director, Institute for Computational Health Sciences

University of California, San Francisco

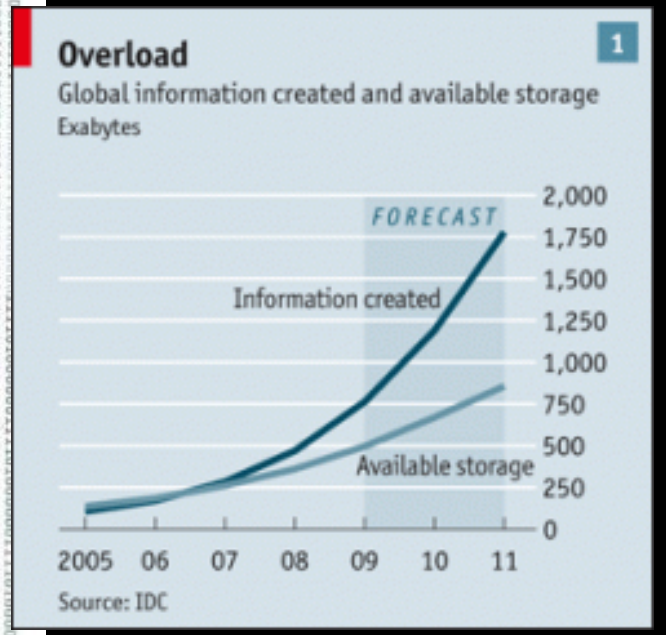
UCSF Health
Redefining possible.

UCSF Benioff Children's Hospital
San Francisco

Conflicts of Interest

- Scientific founder and advisory board membership
  - Genstruct
  - NuMedii
  - Personalis
  - Carmenta
- Honoraria for talks
  - Lilly
  - Pfizer
  - Siemens
  - Bristol Myers Squibb
  - AstraZeneca
  - Roche
  - Genentech
  - Warburg Pincus
- Past or present consultancy
  - Lilly
  - Johnson and Johnson
  - Roche
  - NuMedii
  - Genstruct
  - Tercica
  - Ecoeos
  - Helix
  - Ansh Labs
  - Prevendia
  - Samsung
  - Assay Depot
  - Regeneron
  - Verinata
  - Pathway Diagnostics
  - Geisinger Health
  - Covance
  - Wilson Sonsini Goodrich & Rosati
  - Orrick
  - 10X Genomics
  - Medgenics
  - GNS Healthcare
  - Gerson Lehman Group
  - Coatue Management
- Corporate Relationships
  - Northrop Grumman
  - Aptalis
  - Allergan
  - Astellas
  - Thomson Reuters
  - Intel
  - SAP
  - SV Angel
  - Progenity
  - Illumina
- Speakers' bureau
  - None
- Companies started by students
  - Carmenta
  - Serendipity
  - Stimulomics
  - NunaHealth
  - Praedicat
  - MyTime
  - Flipora
  - Tumbl.in

Kilo
Mega
Giga
Tera
Peta
Exa
Zetta



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

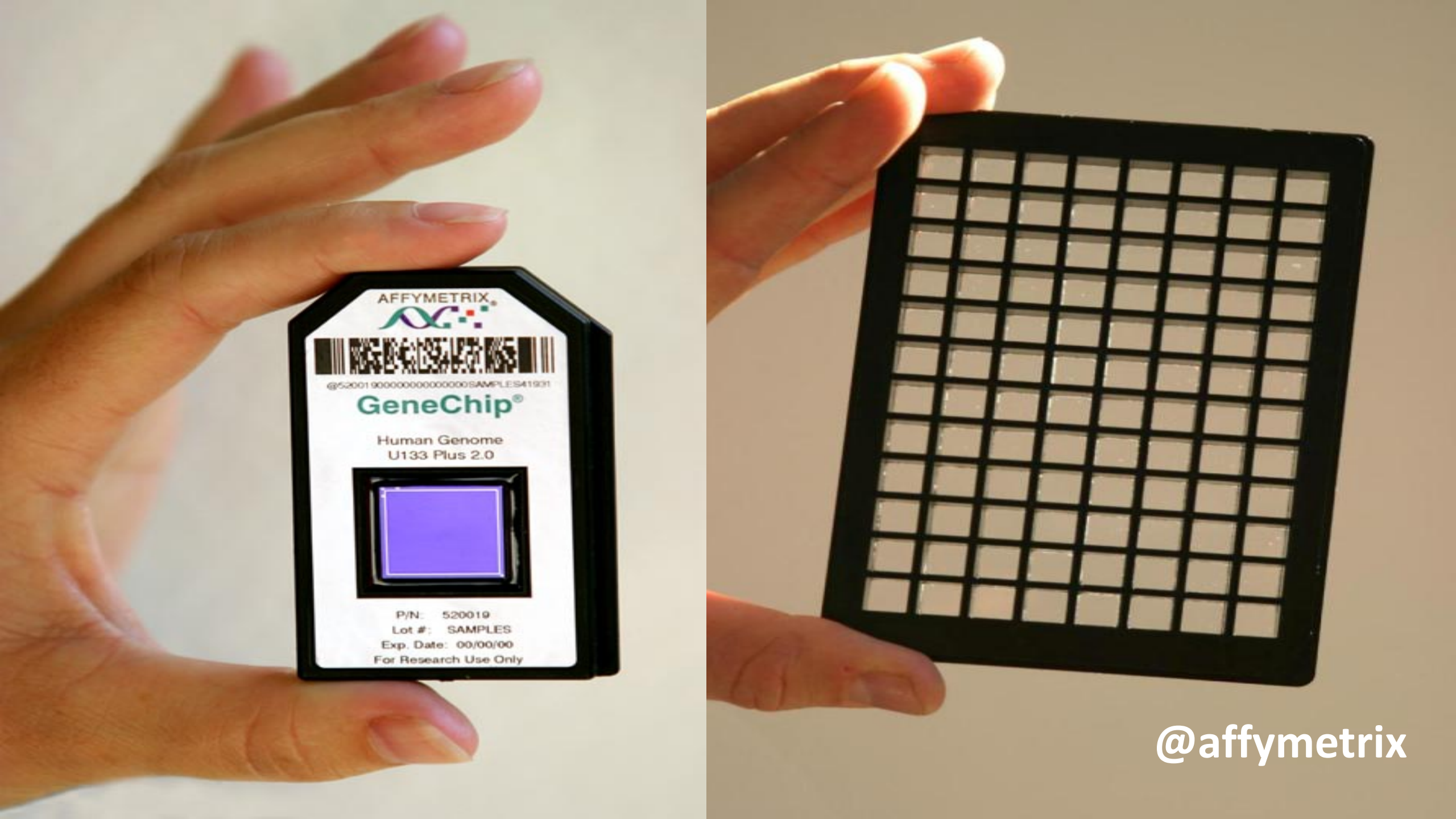
Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30
he was right. But what choice did we have?
from cosmological equations to theories o
behavior, seemed to be able to consistentl
imperfectly, explain the world around us

@chr1sa
bit.ly/endscience


AFFYMETRIX
GeneChip®
Human Genome
U133 Plus 2.0
P/N: 520019
Lot #: SAMPLES
Exp. Date: 00/00/00
For Research Use Only
@affymetrix

DNA microarrays allow researchers to analyse the expression of a huge number of genes simultaneously.

GENOMICS

Gene data to hit milestone

With close to one million gen… …cly accessible repositories, researchers can identify dise… …ter a laboratory.

BY MONYA BAKER

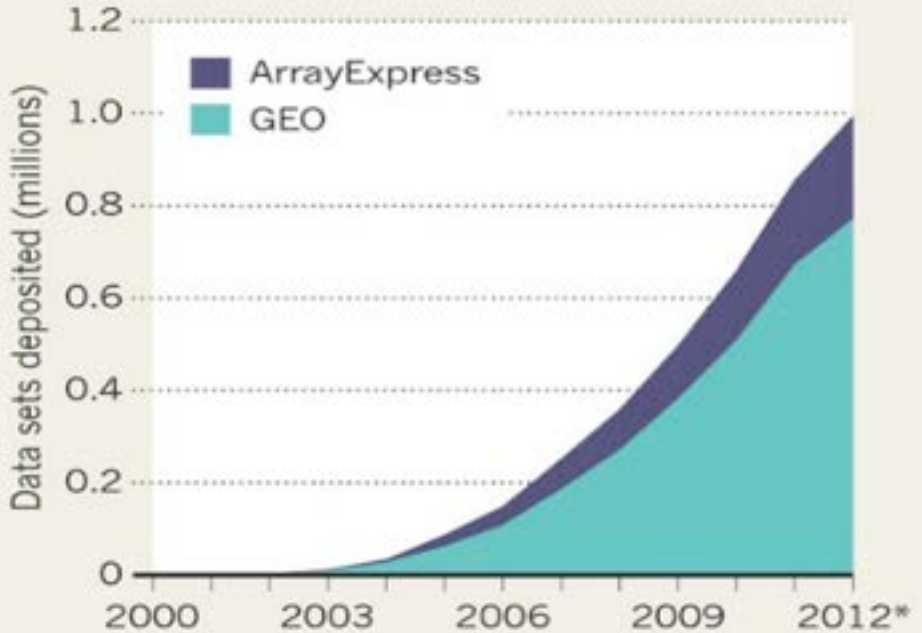
Purvesh Khatri sits in front of an oversize… computer screen, trawling for treasure i… a sea of genetic data. Entering the searc… term 'breast cancer' into a public repositor… called the Gene Expression Omnibus (GEO… the postdoctoral researcher retrieves a list … 1,170 experiments, representing nearly 33,00… samples and a hoard of gene-expression dat… that could reveal previously unseen patterns…

That is exactly the kind of search that le… Khatri's boss, Atul Butte, a bioinformatician … the Stanford School of Medicine in Californi… to identify a new drug target for diabetes. Afte… downloading data from 130 gene-expressio… studies in mice, rats and humans, Butte looke… for genes that were expressed at higher levels i…

…pository at the European Bioinformatics …titute (EBI) in Hinxton, UK. Some time in … next few weeks, the number of deposited …a sets will top one million (see 'Data dump'). The result is an unprecedented resource that …mises to drive down costs and speed up pro- …ss in understanding disease. Gene-sequence …a are already shared extensively, but expres- …n data are more complex and can reveal …ich genes are the most active in, say, liver …sus brain cells, or in diseased versus healthy …ue. And because studies often look at many

DATA DUMP

The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.



bit.ly/genedata

GEO DataSets

GEO DataSets breast cancer

Create alert Advanced

Entry type

DataSets (184)

Series (3,238)

Samples (76,105)

Platforms (49)

Organism

Customize ...

Study type

Expression profiling by array

Methylation profiling by array

Customize ...

Summary 20 per page Sort by Default order

Search results

Items: 1 to 20 of 79576

<< First < Prev

1. MicroRNA-135b overexpression effect on prostate cancer cell line:

Analysis of LNCaP prostate cancer (PCa) cells overexpressing miRNA-135 hours. LNCaP cells express the androgen receptor (AR). MiRNA-135b ove PCa cells results in slower growth compared to AR knockdown. Results pr the basis of this slower growth.

Organism: Homo sapiens

bit.ly/geobrca

GEO DataSets

GEO DataSets breast cancer

Create alert Advanced

Entry type

DataSets (184)

Series (3,238)

Samples (76,105)

Platforms (49)

Organism

Customize ...

Study type

Expression profiling by array

Methylation profiling by array

Customize ...

Summary 20 per page Sort by Default order

Search results

Items: 1 to 20 of 79576

<< First < Prev

1. MicroRNA-135b overexpression effect on prostate cancer cell line:

Analysis of LNCaP prostate cancer (PCa) cells overexpressing miRNA-135 hours. LNCaP cells express the androgen receptor (AR). MiRNA-135b ove PCa cells results in slower growth compared to AR knockdown. Results pr the basis of this slower growth.

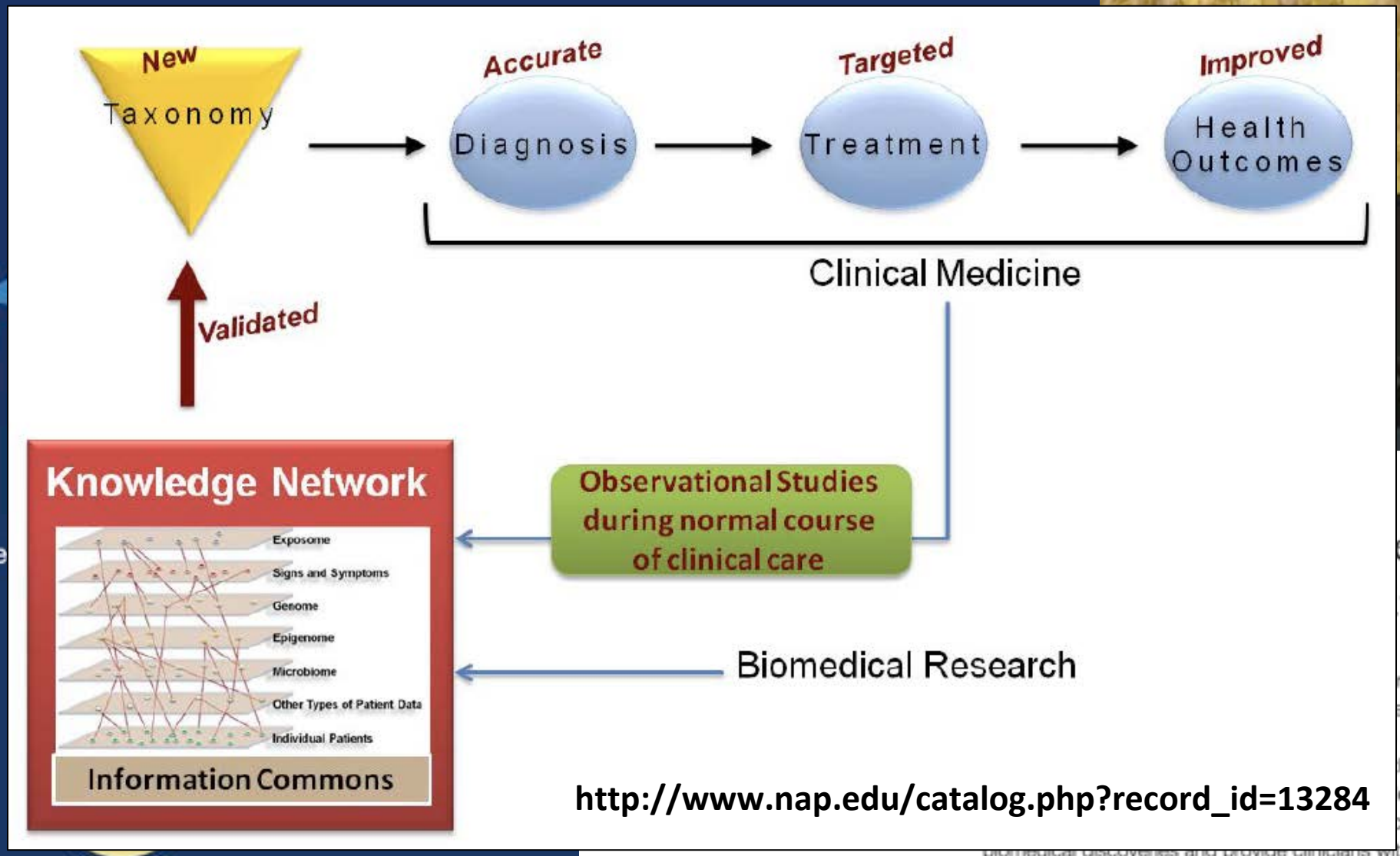
Organism: Homo sapiens

Public big data = retroactive crowd-sourcing

THE PRECISION MEDICINE INITIATIVE

The **time is right** be

Sequencing of the human genome

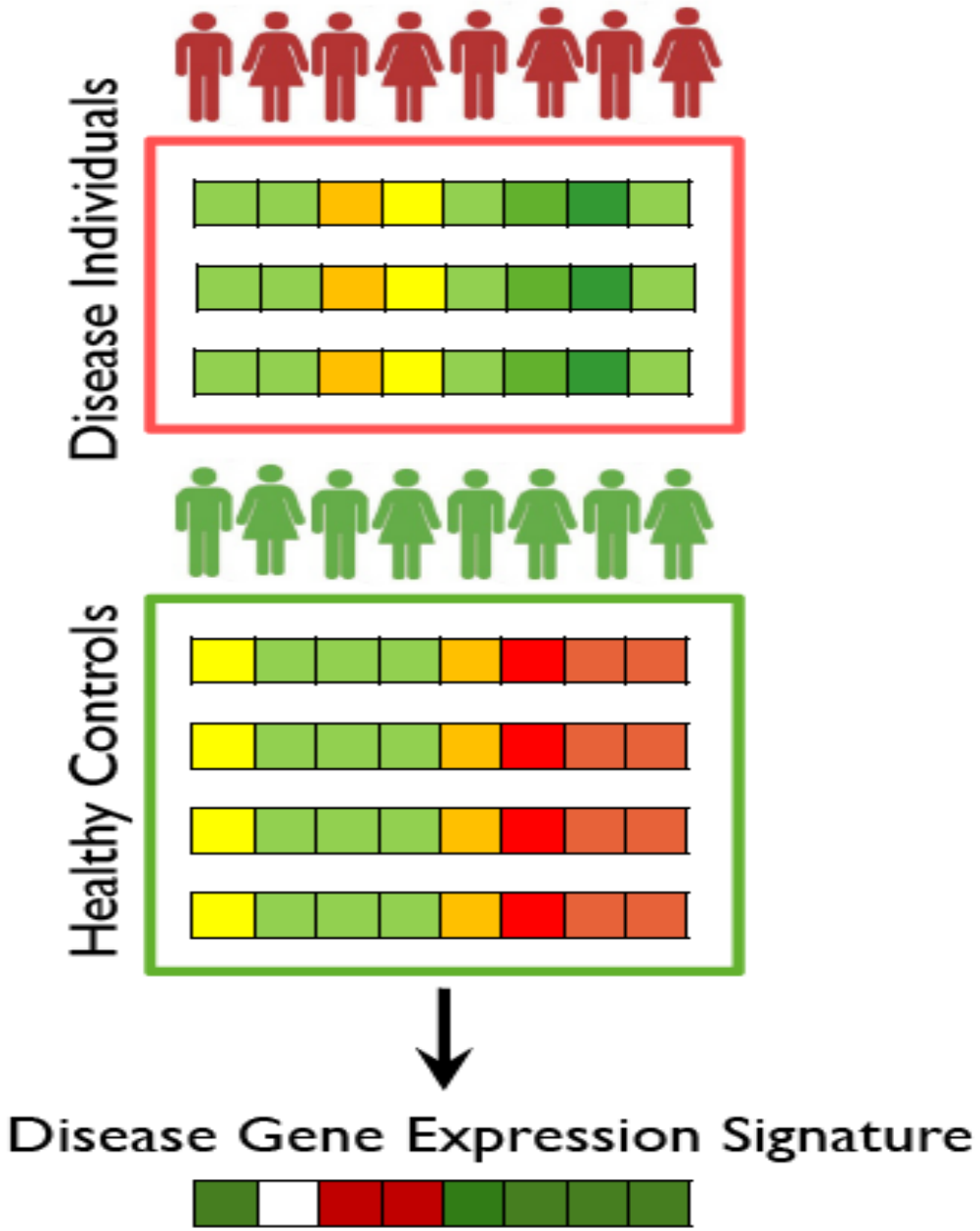


January 30, 2015

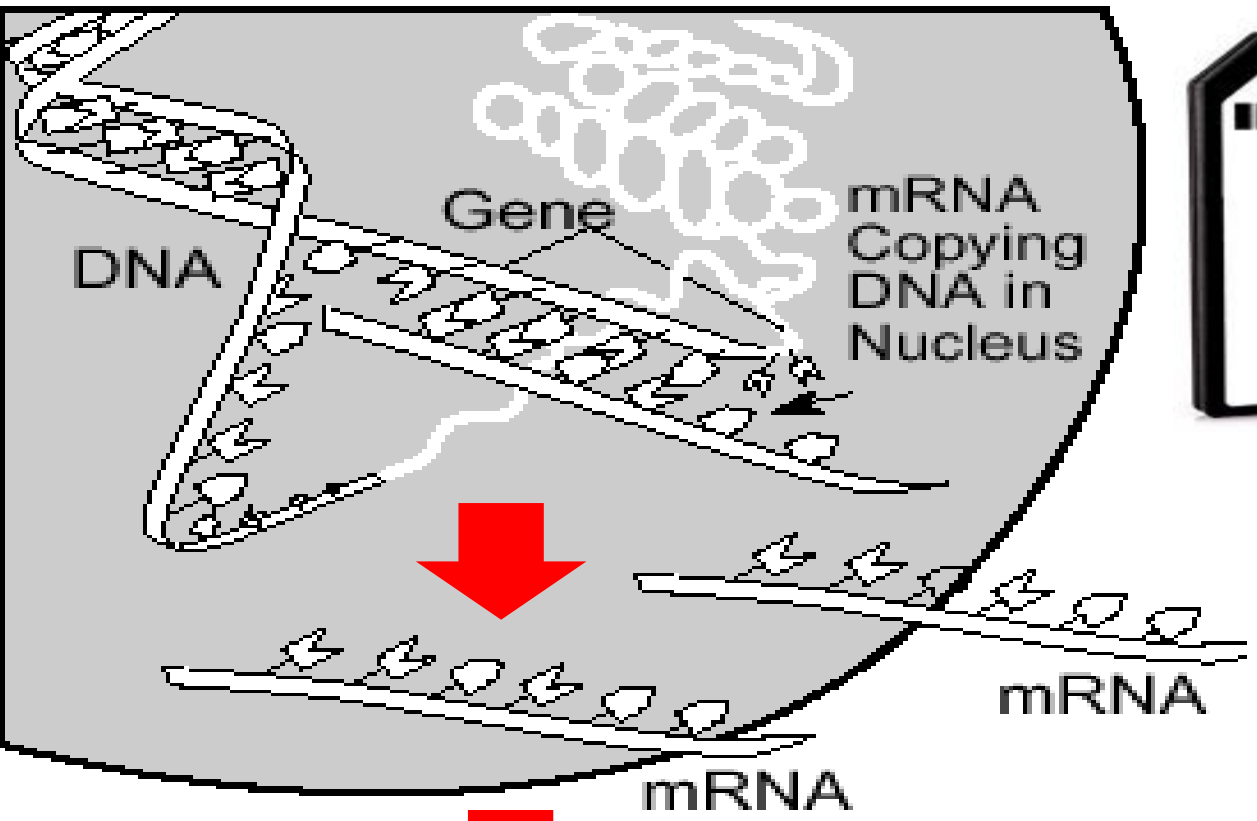
ent Obama's
tiative

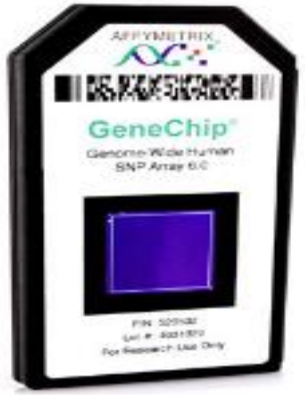
n his State of the Union
tails about the Precision
revolutionize how we
h a $215 million investment
edicine Initiative will pioneer a
omises to accelerate
biomedical discoveries and provide clinicians with new tools, knowledge, and
therapies to select which treatments will work best for which patients.

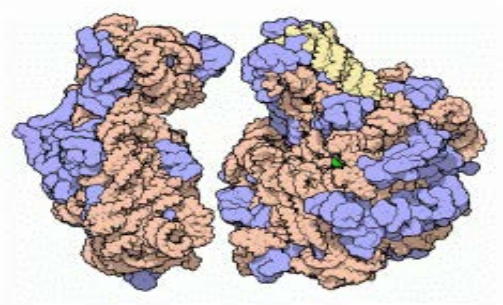
http://www.nap.edu/catalog.php?record_id=13284

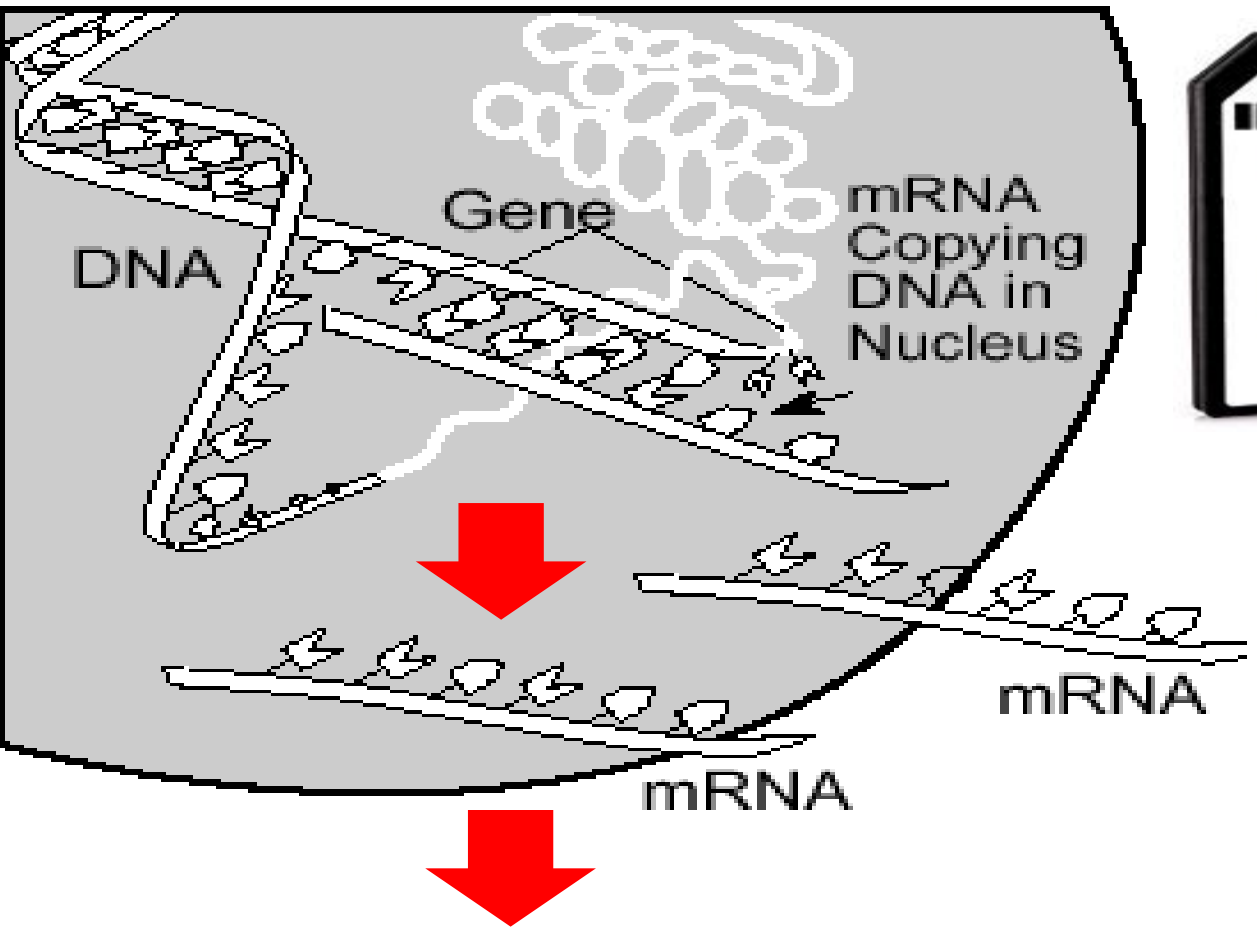


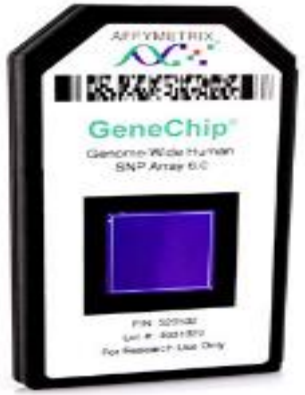
Marina Sirota


DNA
Gene
mRNA
Copying
DNA in
Nucleus
mRNA
mRNA


GeneChip

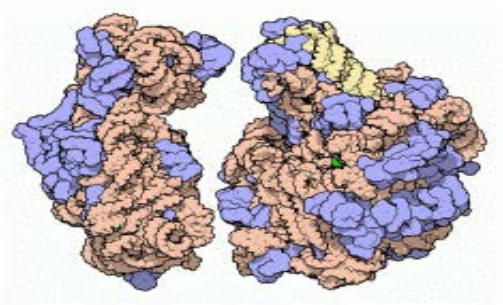


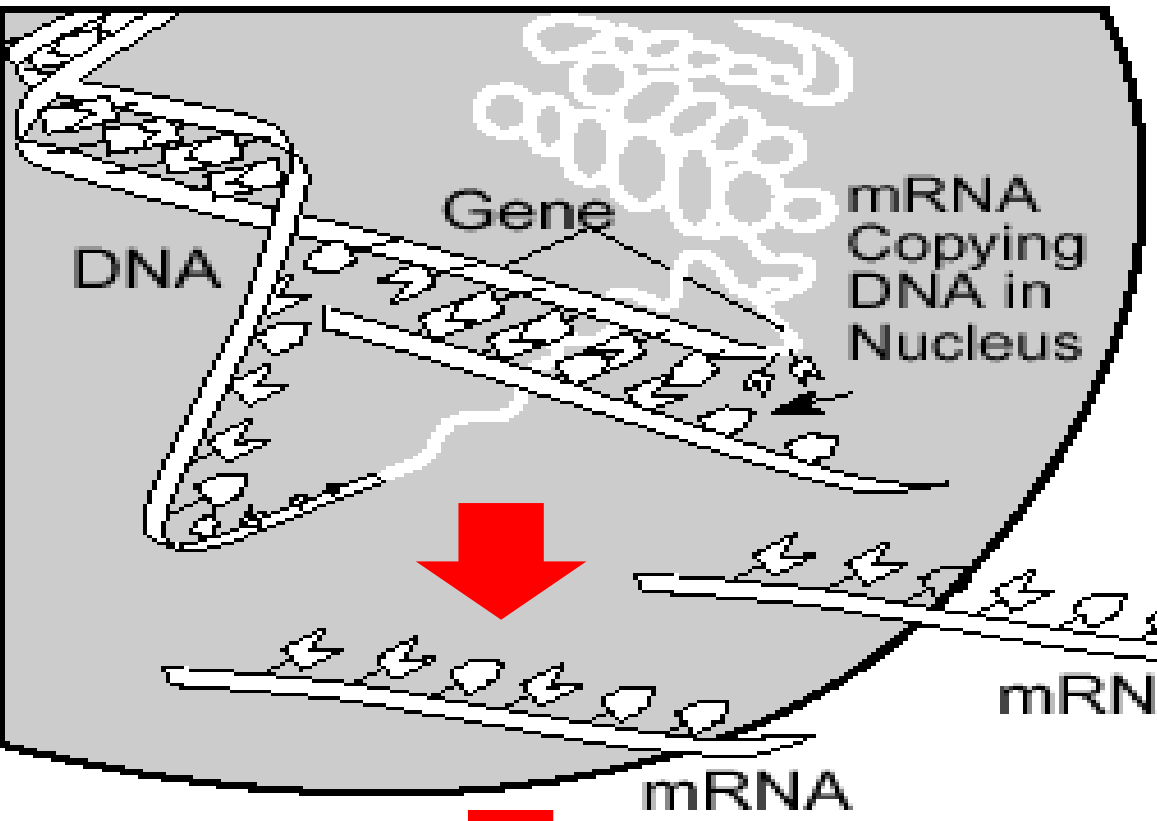




Cancer markers

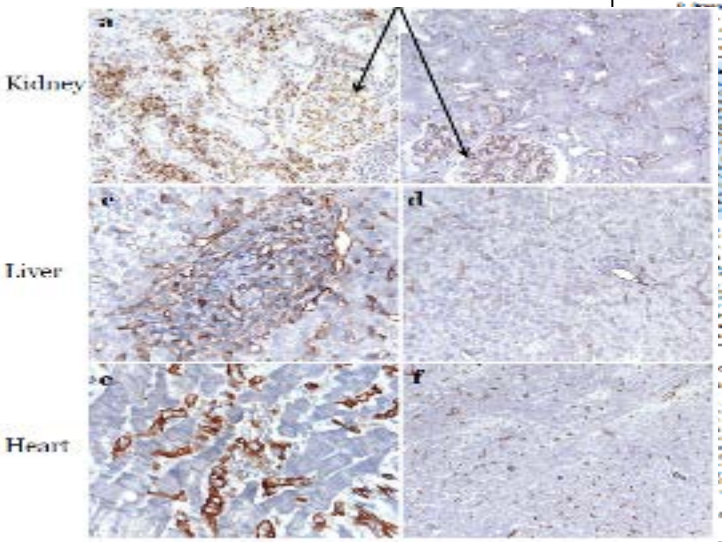
Haeberle H, Dudley JT, …,
Butte AJ, Contag CH. *Neoplasia,* 2012.

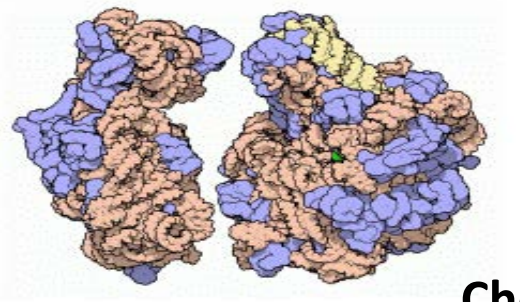




Cancer markers

Transplant Rejection markers





Chen R, …, Butte AJ.
PLoS Computational Biology, 2010.

Preeclampsia: large cause of maternal and fetal death

- Incidence
  - 5-8% of all pregnancies in the U.S. and worldwide
  - 4.1 million births in the U.S. in 2009
  - Up to 300K cases of preeclampsia annually in the U.S.
- Mortality
  - Responsible for 18% of all maternal deaths in the U.S.
  - Maternal death in 56 out of every 100,000 live births in US
  - Neonatal death in 71 out of every 100,000 live births in US
- Cost
  - $20 billion in direct costs in the U.S annually
  - Average hospital stay of 3.5 days

Linda Liu
Bruce Ling
Matt Cooper

Page 1 2 3 4 5 6 .. 11 Showing 1 - 25 of 266 experiments

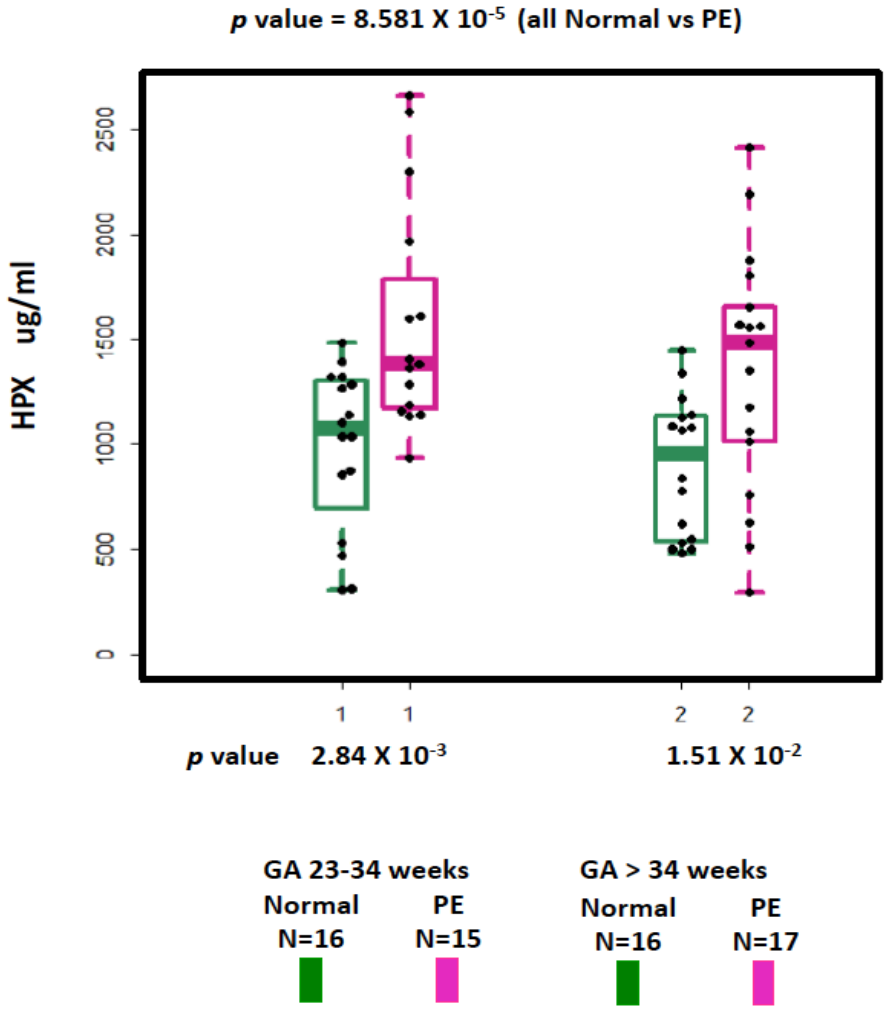
Accession	Title	Type	Organism	Assays	Released
E-GEOD-32472	Oxygen induced complication of prematurity: from experimental data to prevention strategy	transcription profiling by array	Homo sapiens	299	01/11/2011
E-GEOD-27976	Calvarial osteoblast transcriptome analysis identifies genetic targets and extracellular matrix-mediated focal adhesion as potential biomarkers for single-suture craniosynostosis	transcription profiling by array	Homo sapiens	249	04/03/2012
E-GEOD-46510	New whole blood gene expression profile predictve of preterm birth	transcription profiling by array	Homo sapiens	154	15/05/2014
E-GEOD-37210	The application of nonsense-mediated mRNA decay inhibition to the identification of breast cancer susceptibility genes	transcription profiling by array	Homo sapiens	143	11/04/2012
E-TABM-682	Transcription profiling of human decidua basalis to identify pre-eclampsia susceptibility genes	transcription profiling by array	Homo sapiens	104	07/04/2009
E-GEOD-35574	Differentially expressed microRNAs revealed by molecular signatures of Preeclampsia and IUGR in human placenta	transcription profiling by array	Homo sapiens	94	07/02/2012
E-GEOD-41336	Cultured Cyto and Syncytio-trophoblast samples exposed to varying degrees of hypoxia (methylation)	methylation profiling by array	Homo sapiens	90	18/01/2013
E-GEOD-5999	Transcription profiling of human 27 non-	transcription	Homo sapiens	72	07/11/2008

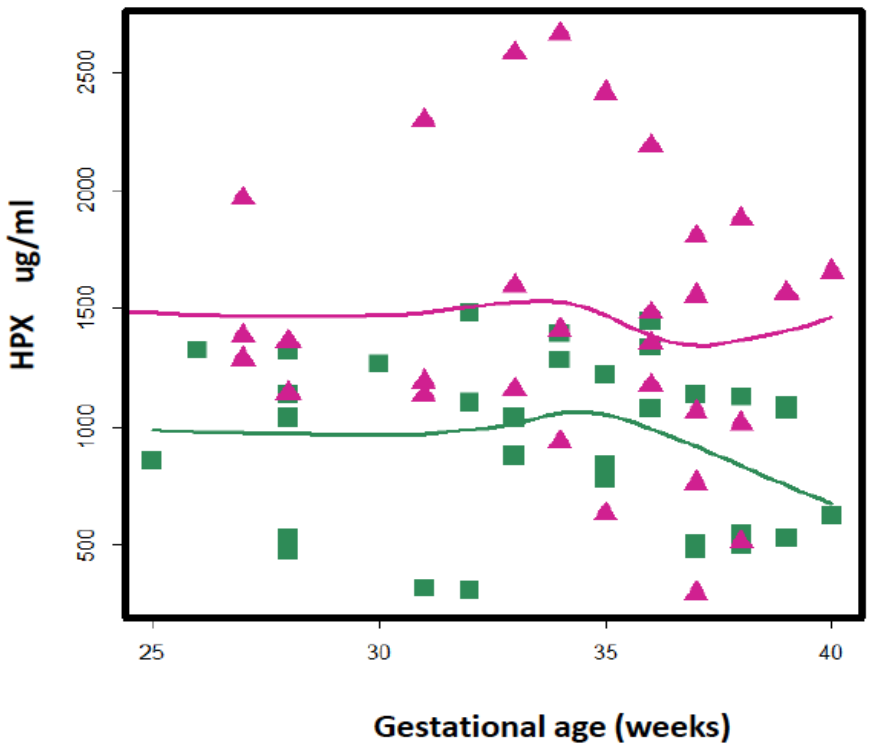
march of dimes
prematurity research center
at STANFORD University School of Medicine

SPARK
AT STANFORD

Linda Liu
Bruce Ling
Matt Cooper

New blood markers for preeclampsia





@MarchofDimes
bit.ly/preeclamp

Need a diagnostic for preeclampsia

Public big data available

March of Dimes Center for Prematurity Research

Data analyzed, diagnostic designed

SPARK grant ($50k)

Life Science Angels, other seed investors ($2 million)

@CarmentaBio
progenity.com
bit.ly/carm_prog

STOCK WATCH Express, Wet Seal, Avago Jump

Carmenta Bioscience Secures Over $2 Million in Oversubscribed Seed Financing

Camille Samuels Accepts Seat on Carmenta Board of Di

BusinessWire

Press Release: Carmenta Bioscience, Inc. – Wed, A

Email Recommend Tweet Share +1

RELATED CONTENT

Carmenta BIOSCIENCE

PALO ALTO, Calif.--(BUSINESS

Carmenta Bioscience, Inc., a priv
company focused on maternal ar
it has secured over $2M in seed f
financing will support development and commercialization of a ne
diagnose and predict preeclampsia in pregnant women. The test is

BusinessWire
A Berkshire Hathaway Company

Progenity Acquires Carmenta Bioscience for Proprietary Preeclampsia Technology; Appoints Matthew Cooper Chief Scientific Officer

April 29, 2015 08:00 AM Eastern Daylight Time

SAN DIEGO--(BUSINESS WIRE)--Progenity, Inc., a provider of complex molecular and specialized diagnostic testing services, today announced the acquisition of Carmenta Bioscience, a leader in preeclampsia diagnostic development. With this acquisition, Progenity continues its mission of helping families prepare for life, through the development of diagnostic tests for preeclampsia. Preeclampsia, a hypertensive disorder of pregnancy, is often difficult to distinguish

Matthew Herper
Forbes Staff

FOLLOW

PHARMA & HEALTHCARE 8/11/2013 @ 11:10AM | 74,127 views

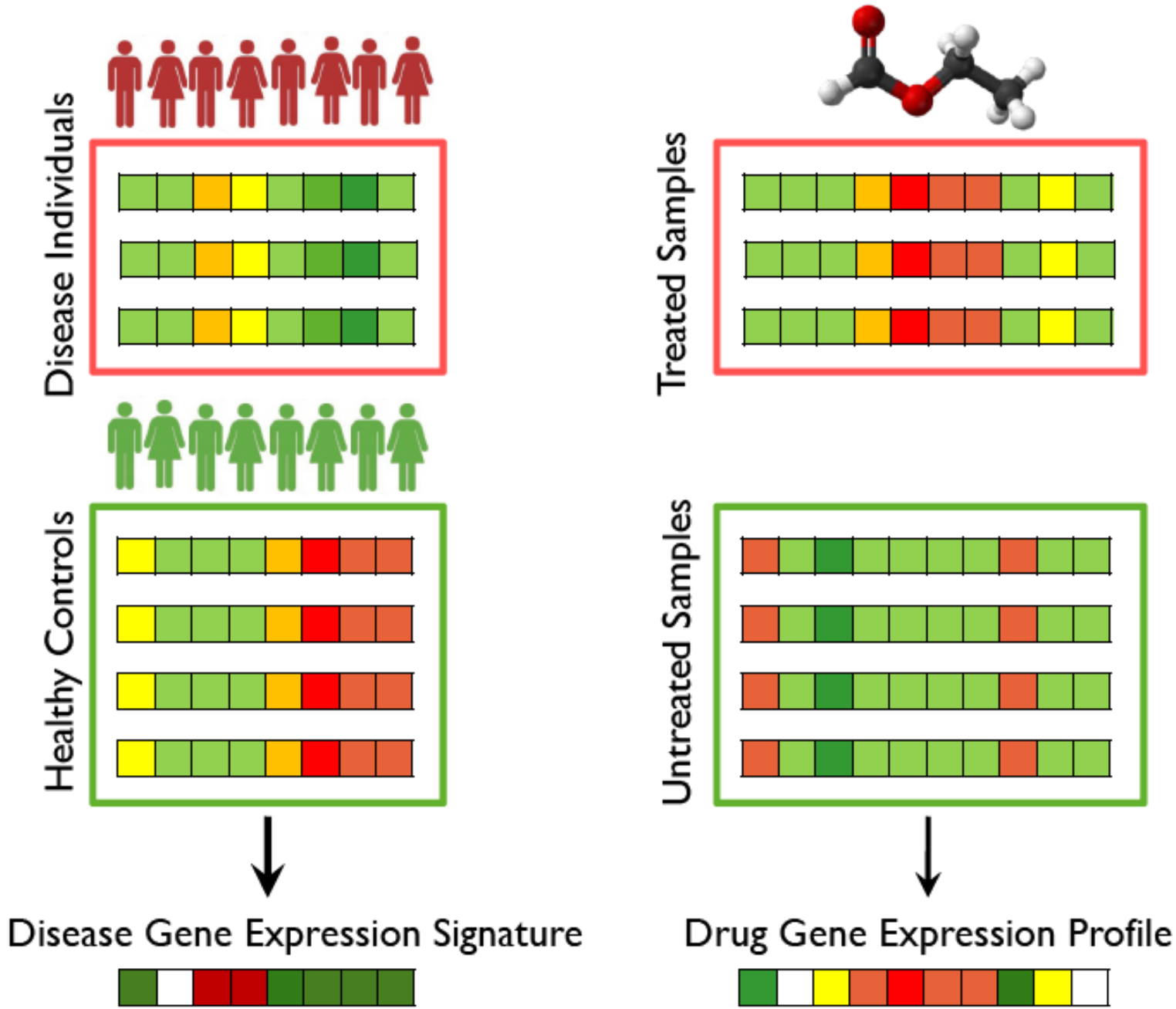
How Much Does Pharmaceutical Innovation Cost? A Look At 100 Companies

+ Comment Now + Follow Comments

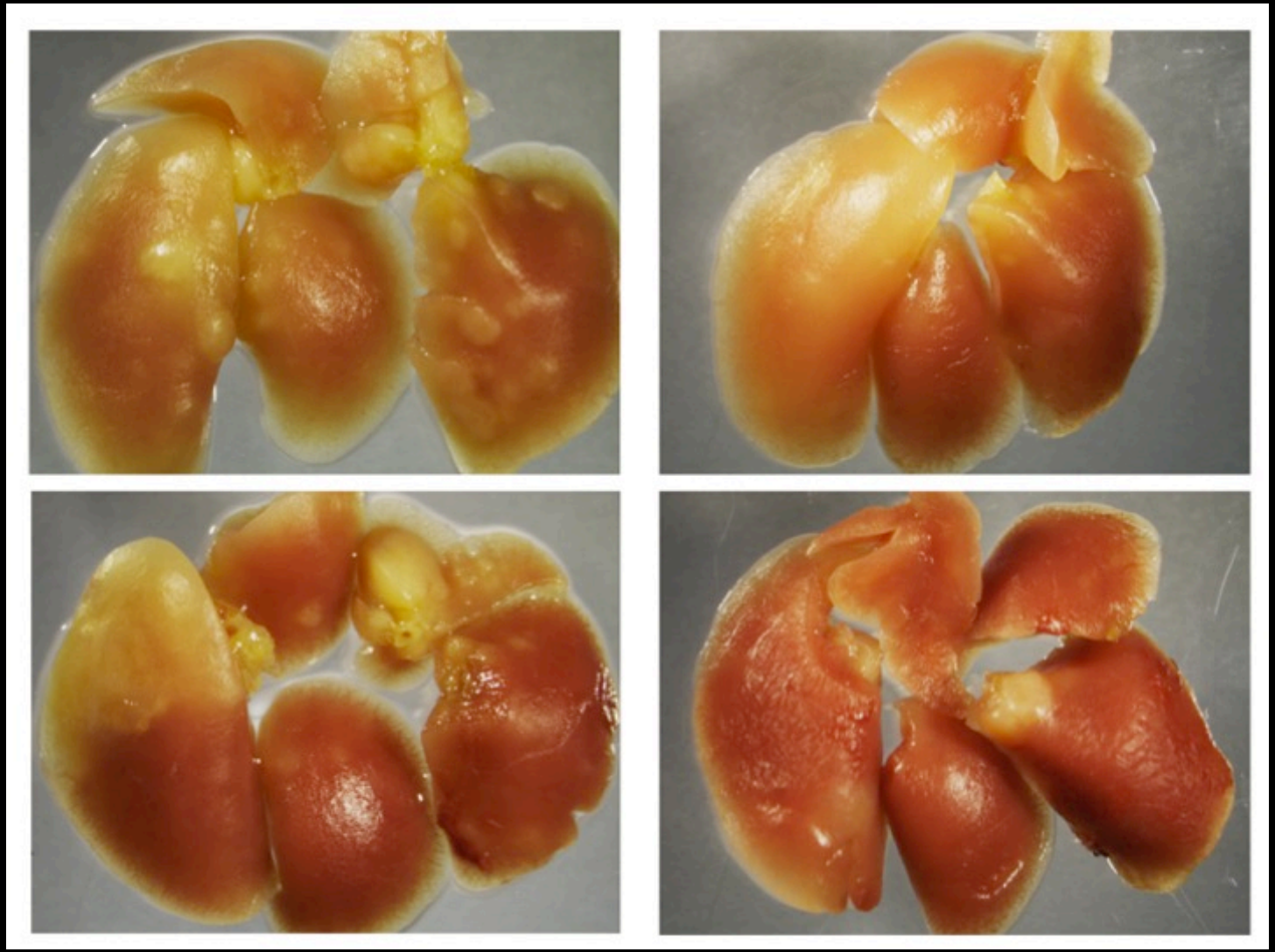
Company	Ticker	Number of drugs approved	R&D Spending Per Drug ($Mil)	Total R&D Spending 1997-2011 ($Mil)
AstraZeneca	AZN	5	11,790.93	58,955
GlaxoSmithKline	GSK	10	8,170.81	81,708
Sanofi	SNY	8	7,909.26	63,274
Roche Holding AG	RHHBY	11	7,803.77	85,841
Pfizer Inc.	PFE	14	7,727.03	108,178
Johnson & Johnson	JNJ	15	5,885.65	88,285
Eli Lilly & Co.	LLY	11	4,577.04	50,347
Abbott Laboratories	ABT	8	4,496.21	35,970
Merck & Co Inc	MRK	16	4,209.99	67,360
Bristol-Myers Squibb Co.	BMY	11	4,152.26	
Novartis AG	NVS	21	3,983.13	
Amgen Inc.	AMGN	9	3,692.14	

Sources: InnoThink Center For Research In Biomedical Innovation; Tho Fundamentals via FactSet Research Systems

@MatthewHerper
bit.ly/newdrug1


Disease Individuals
Healthy Controls
Disease Gene Expression Signature
Treated Samples
Untreated Samples
Drug Gene Expression Profile

Psychiatric Drug Imipramine Shows Significant Activity Against Small Cell Lung Cancer



Vehicle control

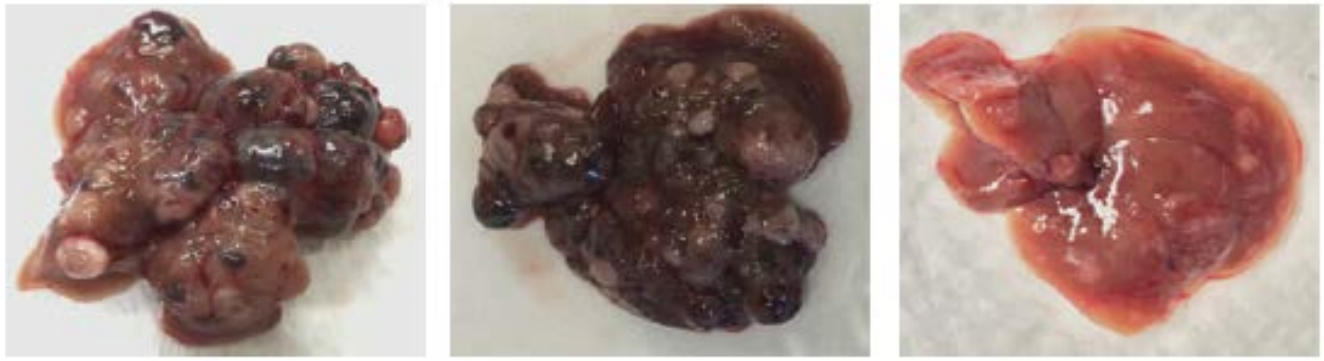
Imipramine

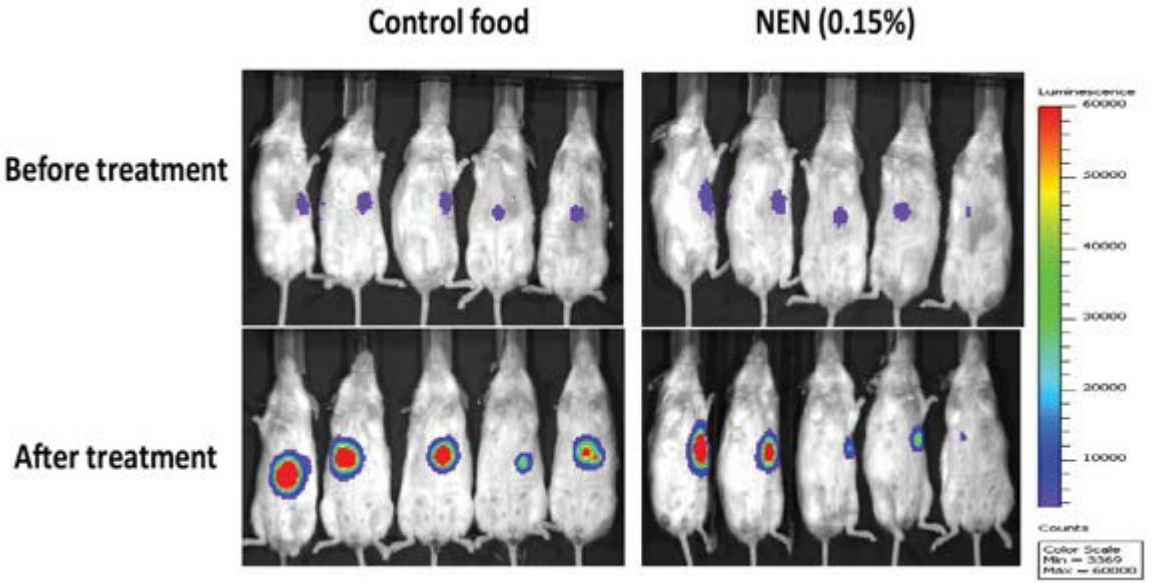
p53/Rb/p130 triple knockout model of SCLC

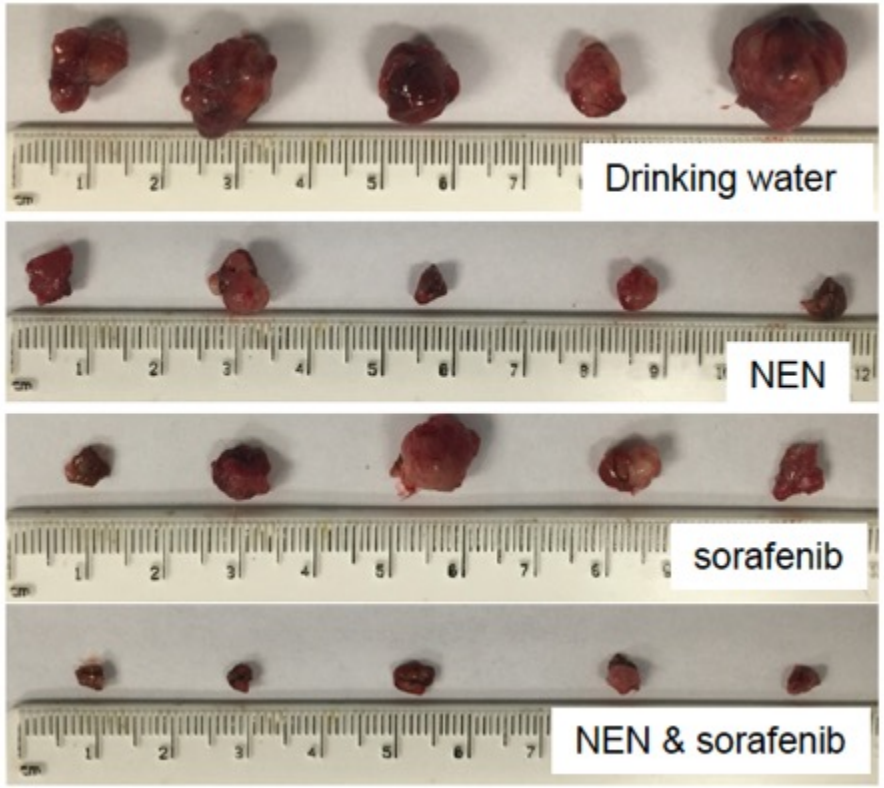
Mice dosed after tumor formation

Joel Dudley
Nadine Jahchan
Julien Sage
Alejandro Sweet-Cordero
Joel Neal
@NuMedii

***Cancer Discovery* 2013, 3:1.**







Bin Chen
Wei Wei
Li Ma
Bin Yang
Mei-Sze Chua
Samuel So
***Gastroenterology*, 2017**

Need more drugs for more diseases

Public big data available

NIH funding

Data analyzed, method designed

Company launched, ARRA, StartX, Stanford license, first deal

Claremont Creek, Lightspeed ($3.5 million)

@NuMedii

NuMedii
Translating Big Data into new medicines

Venture capital

'Digital drug development' company NuMedii snags $3.5 million

Ron Leuty
Reporter-
San Francisco Business Times
Email | Twitter | Google+ | Twitter

NuMedii Inc., the Palo Alto startup looking to convert pages of drug safety data into faster drug-development times, lined up $3.5 million in a Series A round.

Enlarge

NuMedii CEO Gini Deshpande: Tapping old data drugs.

The oversubscribed round was led by Claremont Creek Ventures and Lightspeed Ve Partners and included Life Science Angels and others.

NuMedii's data-into-gold approach rolls a wide range of data — from public scientific data bases and other sources — into an algorithm to predict if a compound will trans

THE WALL STREET JOURNAL. | U.S.

TOP STORIES IN U.S. 1 of 12

VA Scandal Is Shinseki's Latest Battle

2 of 12

U.S. Economy Contracted Last Quarter

U.S. NEWS

Researchers Show Gains in Finding Reus

Email Print

By AMY DOCKSER

August 18, 2011

In a bit of high-tec already-approved combat.

The scientists ha have benefits in t

FierceBiotechIT

Topics: R&D

Allergan taps NuMedii's digit platform for psoriasis R&D

October 5, 2015 | By Nick Paul Taylor

SHARE

NuMedii has landed a deal that could vali discovery. Allergan ($AGN) is the compa tments fo

Astellas hooks up with NuMedii to continue drug repurposing deal drive

January 15, 2016 | By Nick Paul Taylor

SHARE

Email

NuMedii, Inc. Announces New Partnership To Discover An Advance New Treatments For Idiopathic Pulmonary Fibros

immport.org

Sanchita Bhattacharya
Jeff Wiser

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

ImmPort is funded by the NIH, NIAID and DAIT in support of the NIH mission to share data with the public. Data shared through ImmPort has been provided by NIH-funded programs, other research organizations and individual scientists ensuring these discoveries will be the foundation of future research.

Welcome to Immport.org our new Beta web site currently in user review

Data Summary - Studies: 222, Subjects: 37140, Experiments: 1011

more >

Private Data

NIH National Institute of Allergy and Infectious Diseases

Data Analysis

- Analysis Workflow
- Automated Clustering
- Tutorials

NIH National Institute of Allergy and Infectious Diseases

march of dimes

BILL & MELINDA GATES foundation

The next big open data: clinical trials

Download 100+ studies today

Drug repositioning, new patient subsets, digital comparative effectiveness, more!

NORTHROP GRUMMAN

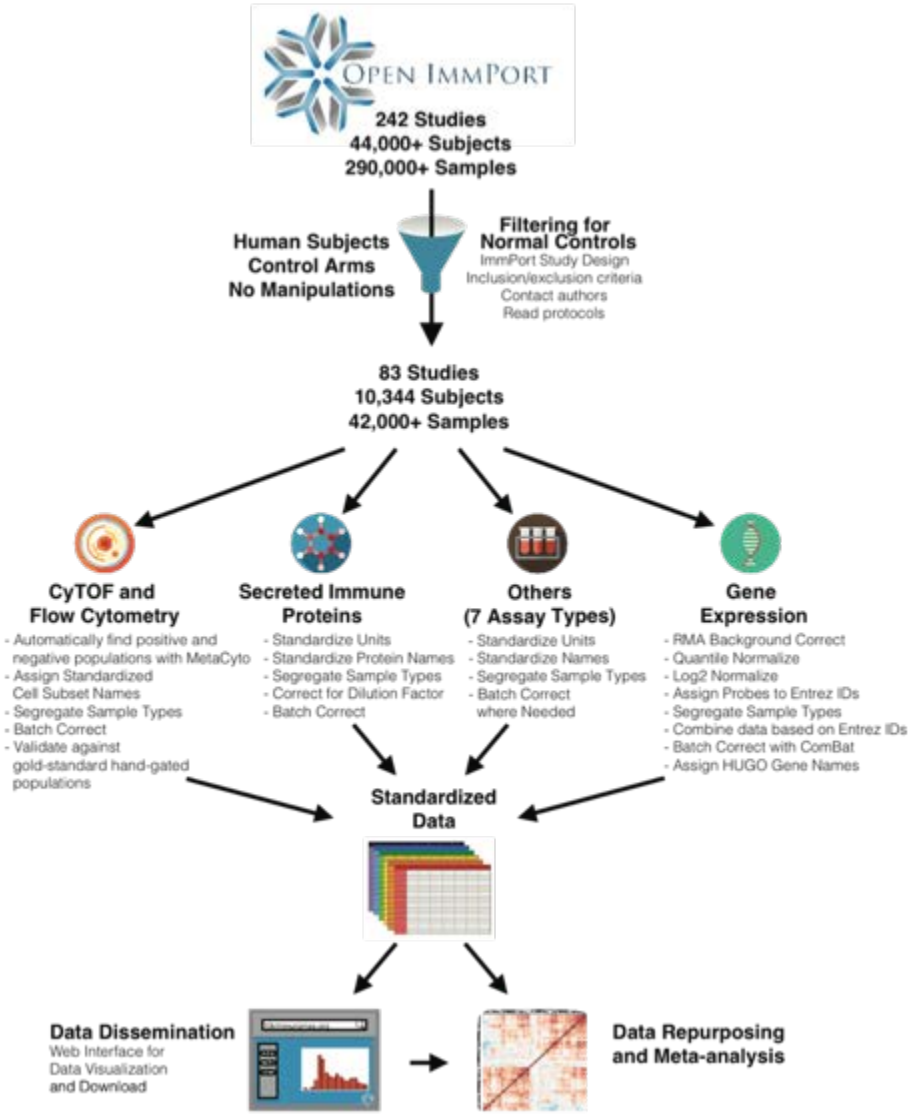
UCSF

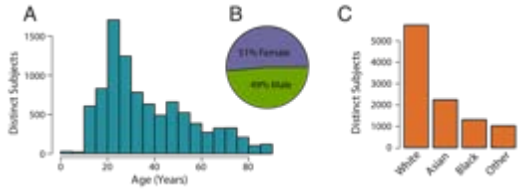
LELAND STANFORD JUNIOR UNIVERSITY 1891

Technion Israel Institute of Technology

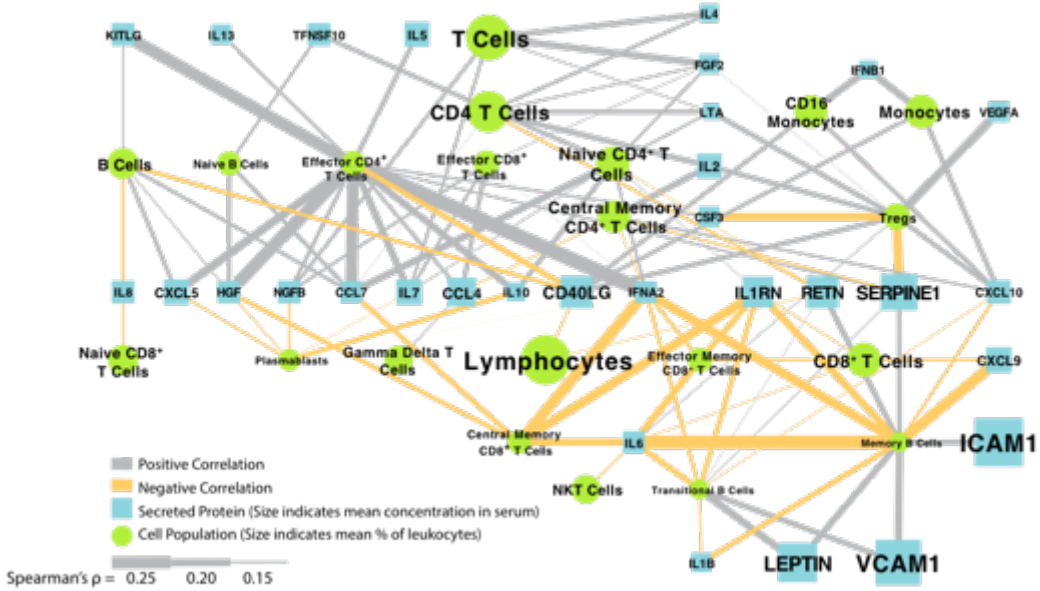
University at Buffalo
The State University of New York

The 10,000 Immunome Project: From the control groups of 242 manually curated experiments





Data available in the 10,000 Immunomes Project	
Total Samples	42117
Total Distinct Subjects	10344
MEASUREMENT	**SUBJECTS**
Secreted Proteins	4835
ELISA	4035
Multiplex ELISA	1286
Virus Titer	3609
Virus Neutralization Titer	2265
HAI Titer	1344
Clinical Lab Tests	2639
Complete Blood Count	1684
Comprehensive Metabolic Panel	664
Fasting Lipid Profile	664
Questionnaire	1422
Cytometry	1415
Flow Cytometry (PBMC)	907
CyTOF (PBMC)	583
Flow Cytometry (Whole Blood)	164
HLA Type	1093
Gene Expression Array	476
Whole Blood	311
PBMC	165



Kelly Zalocusky
Sanchita Bhattacharya
@ImmPortDB

bioRxiv bit.ly/10kimmu
http://10kimmunomes.org/

UCSF Institute for Computational Health Sciences

San Francisco Chronicle

Zuckerberg, Chan give UCSF $10 million for health data research

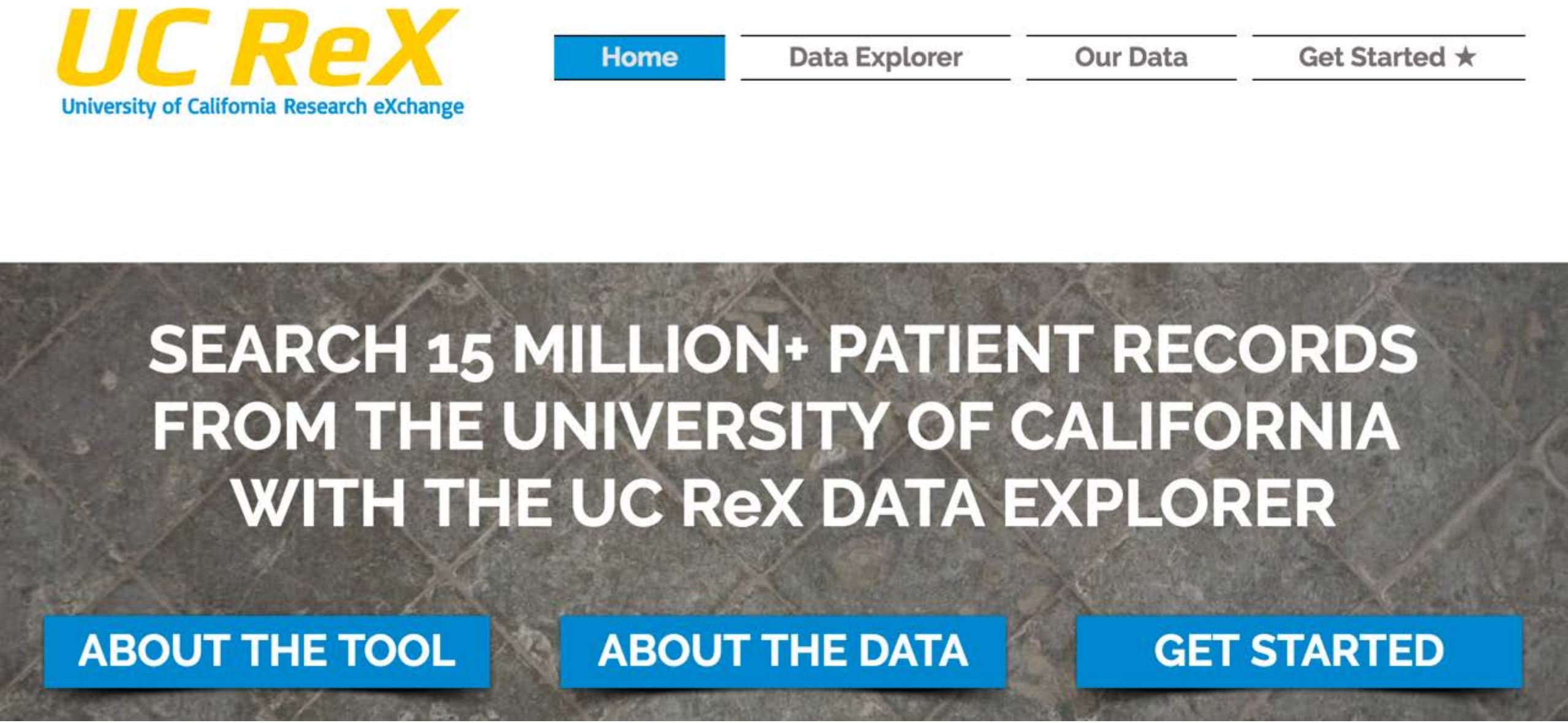
By Catherine Ho, San Francisco Chronicle | July 28, 2017

UCSF Institute for Computational Health Sciences

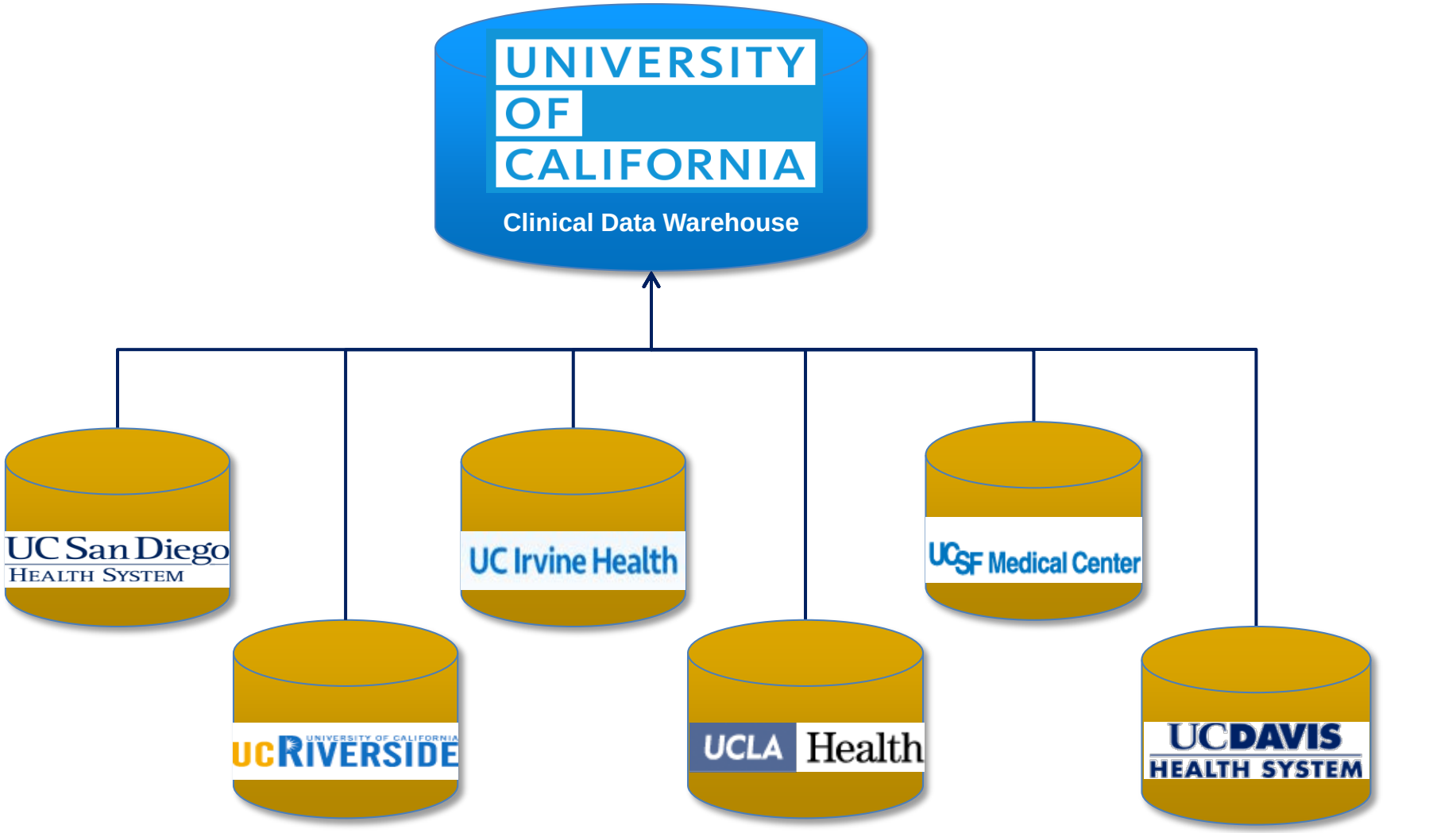
Build the strongest team in the world in biomedical computation and health data analytics

- Academic affinity home for faculty and staff
- Research and development (and spin out technologies)
- Develop new educational plans
- Bring the best new computational and informatics faculty members to UCSF
- Organize infrastructure and operations
- Build and use our new data assets for precision medicine

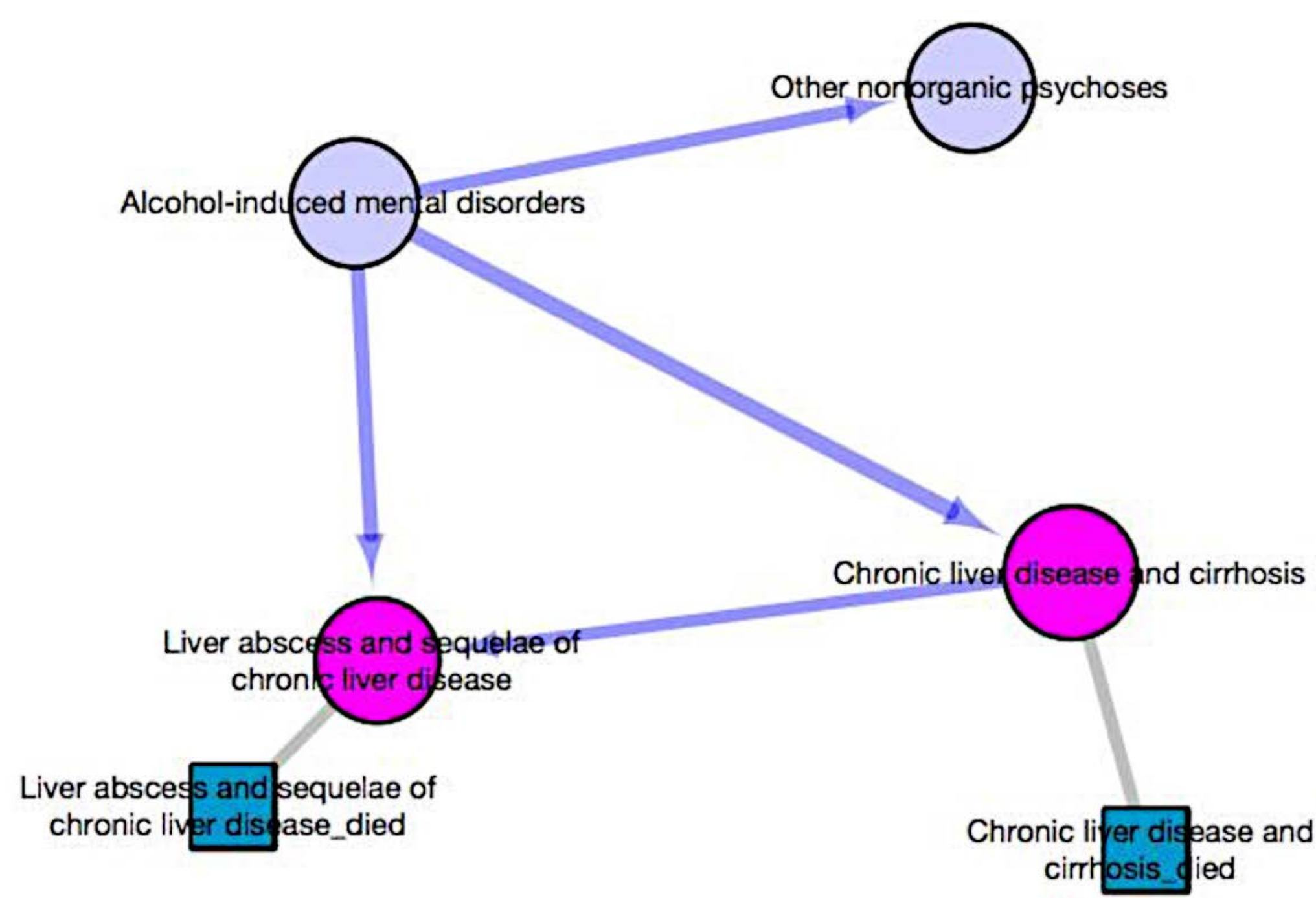
The next big data: clinical data



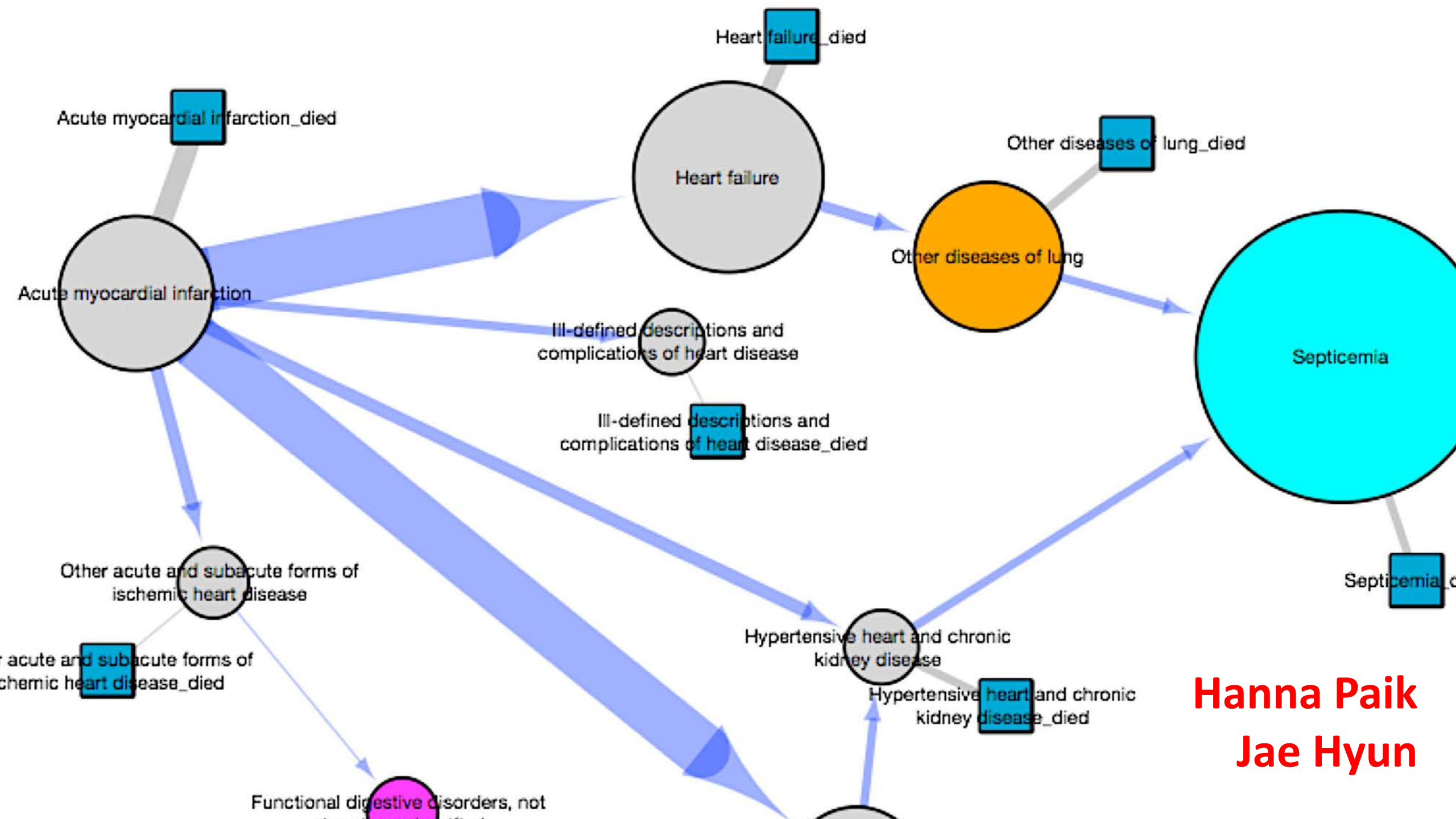
Combining healthcare data from across the six University of California medical schools and systems

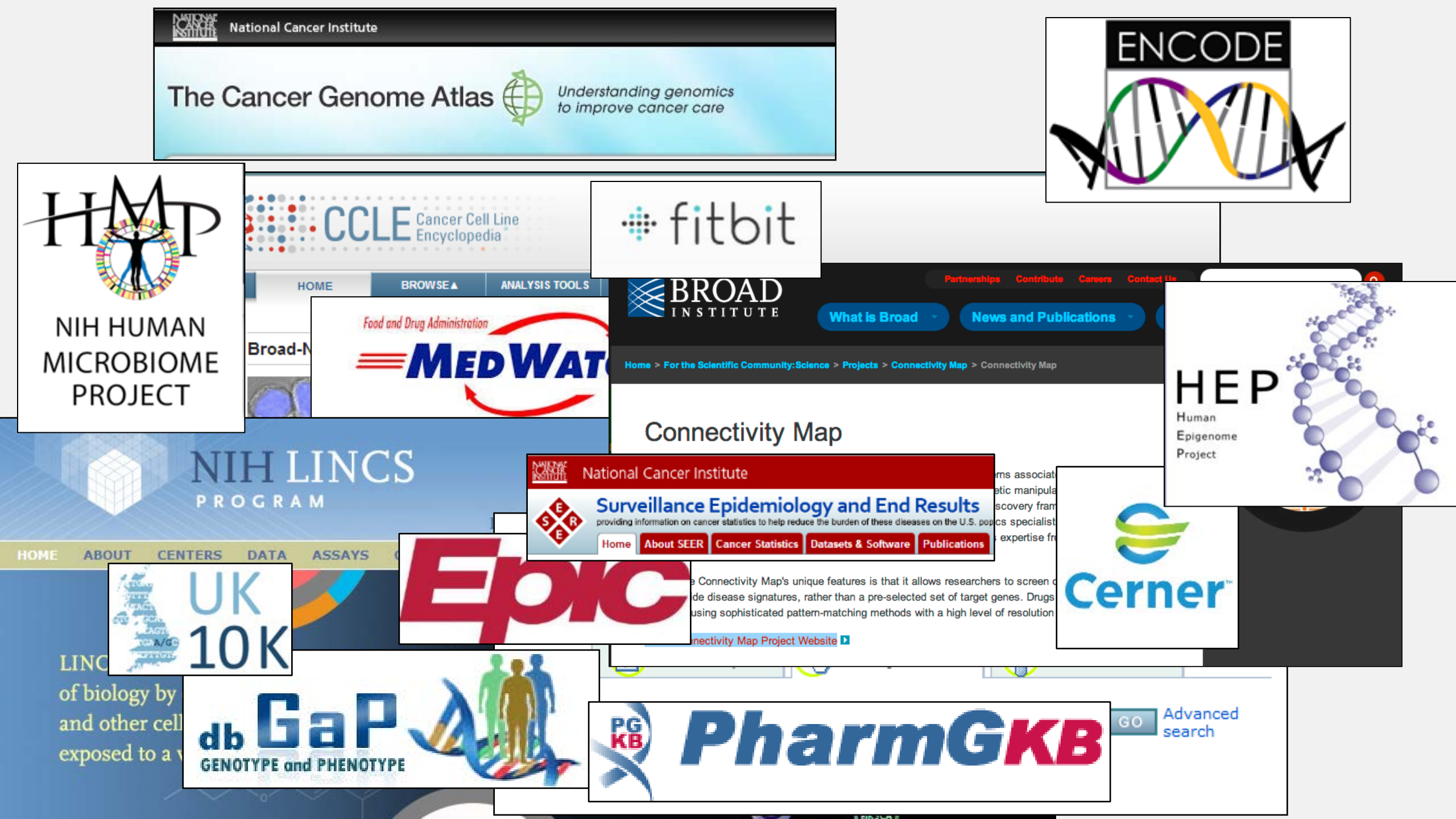


A Big UC Healthcare Data Analytics Platform



Hanna Paik
Jae Hyun


Heart failure_died
Acute myocardial infarction_died
Heart failure
Other diseases of lung_died
Other diseases of lung
Acute myocardial infarction
Ill-defined descriptions and complications of heart disease
Septicemia
Ill-defined descriptions and complications of heart disease_died
Other acute and subacute forms of ischemic heart disease
Hypertensive heart and chronic kidney disease
Hypertensive heart and chronic kidney disease_died
Functional digestive disorders, not
Hanna Paik
Jae Hyun


National Cancer Institute
The Cancer Genome Atlas
Understanding genomics to improve cancer care
ENCODE
HMP
NIH HUMAN MICROBIOME PROJECT
CCLE Cancer Cell Line Encyclopedia
HOME
BROWSE
ANALYSIS TOOLS
fitbit
Food and Drug Administration
MEDWAT
BROAD INSTITUTE
Partnerships
Contribute
Careers
Contact Us
What is Broad
News and Publications
Home > For the Scientific Community:Science > Projects > Connectivity Map > Connectivity Map
Connectivity Map
HEP
Human Epigenome Project
NIH LINCS PROGRAM
HOME
ABOUT
CENTERS
DATA
ASSAYS
National Cancer Institute
Surveillance Epidemiology and End Results
providing information on cancer statistics to help reduce the burden of these diseases on the U.S. pop
Home
About SEER
Cancer Statistics
Datasets & Software
Publications
Epic
Cerner
UK 10K
dbGaP
GENOTYPE and PHENOTYPE
PharmGKB
GO
Advanced search

Take home points:

- Plenty of high-quality data already available: some public, some private
- Don’t wait for perfection; data always getting better
- Use and intersect data to ask new questions, to innovative new diagnostics and drugs
- Academia and industry are compatible: the science can and will continue in industry

UC Clinical Data Warehouse Team

Executive Team
- Atul Butte
- Joe Bengfort
- Michael Pfeffer
- Tom Andriola
- Chris Longhurst

Steering Committee
- Irfan Chaudhry
- Mohammed Mahbouba
- Lisa Dahm
- David Dobbs
- Kent Andersen
- Ralph James
- Jennifer Holland
- Eugene Lee

ETL Team
- Albert Dugan
- Tony Choe
- Michael Sweeney
- Timothy Satterwhite
- Ayan Patel
- Niranjan Wagle
- Ralph James
- Joseph Dalton

Data Harmonization
- Dana Ludwig
- Daniella Meeker

Data Quality
- Momeena Ali
- Jodie Nygaard

Epic
- Kevin Ames
- Ben Jenkins
- Steve Gesualdo

Business Analyst
- Ankeeta Shukla

Hardware
- Sandeep Chandra
- Jeff Love
- Scott Bailey
- Kwong Law
- Pallav Saxena

Support
- Jack Stobo
- Michael Blum
- Sam Hawgood

Collaborators

- Jeff Wiser, Patrick Dunn, Mike Atassi / Northrop Grumman
- Ashley Xia and Quan Chen / NIAID
- Takashi Kadowaki, Momoko Horikoshi, Kazuo Hara, Hiroshi Ohtsu / U Tokyo
- Kyoko Toda, Satoru Yamada, Junichiro Irie / Kitasato Univ and Hospital
- Shiro Maeda / RIKEN
- Jeff Olgin / Cardiology
- Alejandro Sweet-Cordero, Julien Sage / Pediatric Oncology
- Mark Davis, C. Garrison Fathman / Immunology
- Russ Altman, Steve Quake / Bioengineering
- Euan Ashley, Joseph Wu, Tom Quertermous / Cardiology
- Mike Snyder, Carlos Bustamante, Anne Brunet / Genetics
- Jay Pasricha / Gastroenterology
- Rob Tibshirani, Brad Efron / Statistics
- Hannah Valantine, Kiran Khush/ Cardiology
- Ken Weinberg / Pediatric Stem Cell Therapeutics
- Mark Musen, Nigam Shah / National Center for Biomedical Ontology
- Minnie Sarwal / Nephrology
- David Miklos / Oncology

Support

- University of California, San Francisco
- NIH: NIAID, NLM, NIGMS, NCI, NHLBI, OD; NIDDK, NHGRI, NIA, NCATS
- March of Dimes
- Juvenile Diabetes Research Foundation
- Hewlett Packard
- Howard Hughes Medical Institute
- California Institute for Regenerative Medicine
- Luke Evnin and Deann Wright (Scleroderma Research Foundation)
- Clayville Research Fund
- PhRMA Foundation
- Stanford Cancer Center, Bio-X, SPARK

- Tarangini Deshpande
- Kimayani Butte
- Sam Hawgood
- Keith Yamamoto
- Isaac Kohane

Admin and Tech Staff

- Mary Lyall
- Mounira Kenaani
- Kevin Kaier
- Boris Oskotsky
- Mae Moredo
- Ada Chen