# Novel Structures (and Non-Structures) to Facilitate Translational Research

## Integrating layers of omics data models and compute spaces needed to build a "Knowledge Expert"

Stephen Friend MD PhD

Sage Bionetworks (Non-Profit Organization)
Seattle/ Beijing/ Amsterdam

MIT/Whitehead
October 10th, 2011

Why not use data intensive science
to build models of disease

Organizational Structures and Tools

How not What

Six Pilots

Opportunities

Alzheimer's

Diabetes

Treating Symptoms v.s. Modifying Diseases

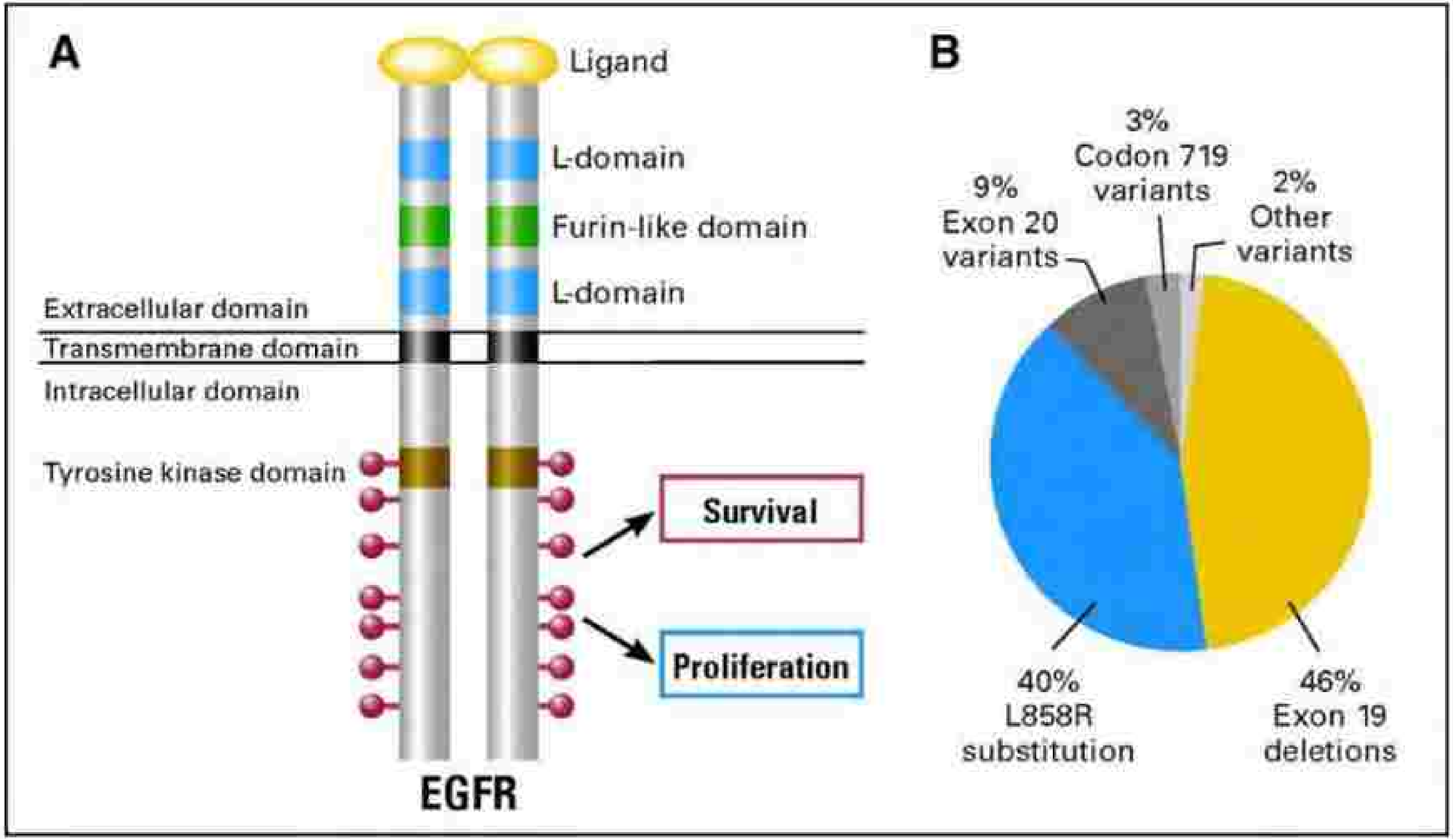
Cancer

Obesity

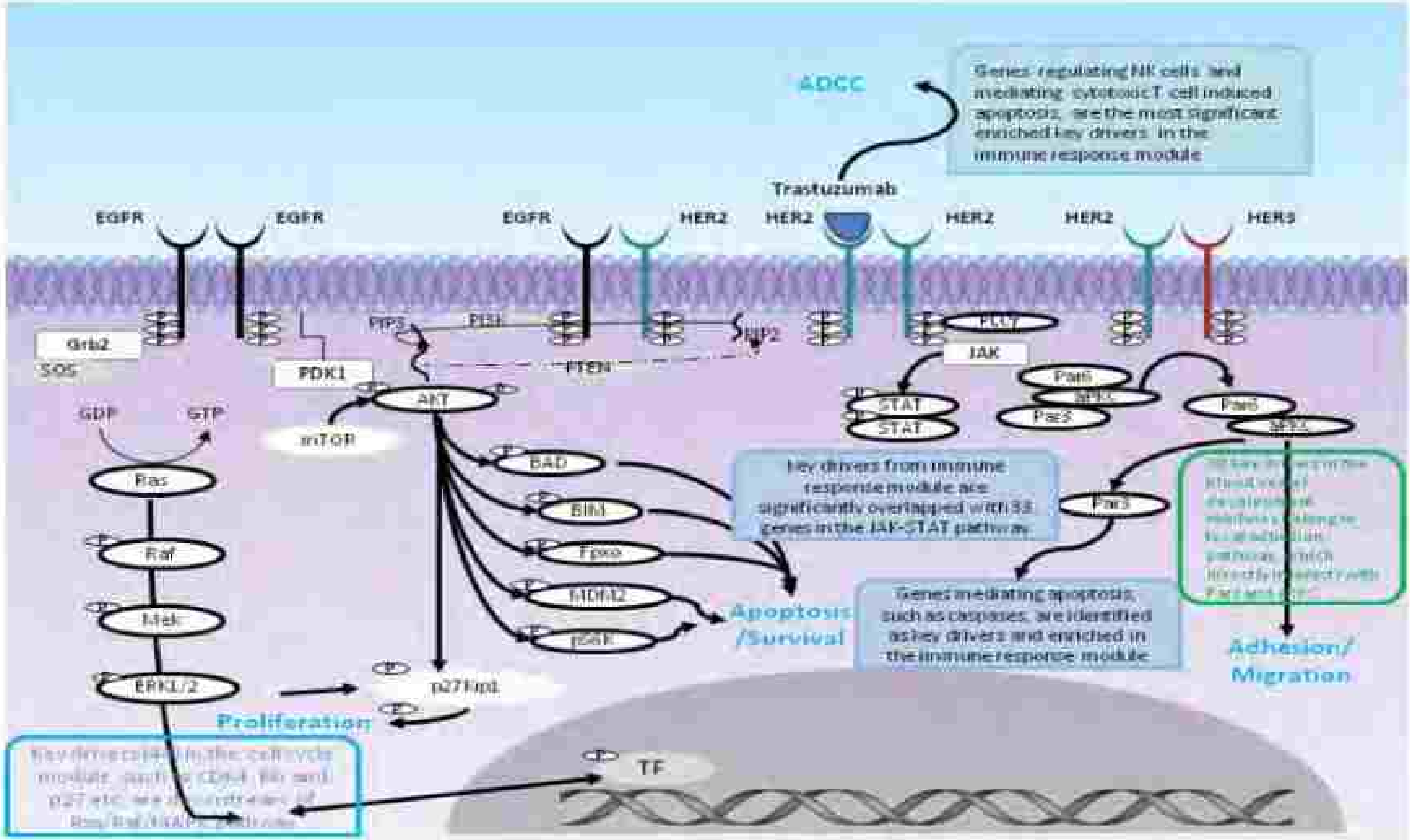
Will it work for me?

Biomarkers?

# Personalized Medicine 101:
# Capturing Single bases pair mutations = ID of responders



# Reality: Overlapping Pathways



Genes regulating NK cells and mediating cytotoxic T cell induced apoptosis, are the most significant enriched key drivers in the immune response module

Key drivers from immune response module are significantly overlapped with 33 genes in the JAK-STAT pathway

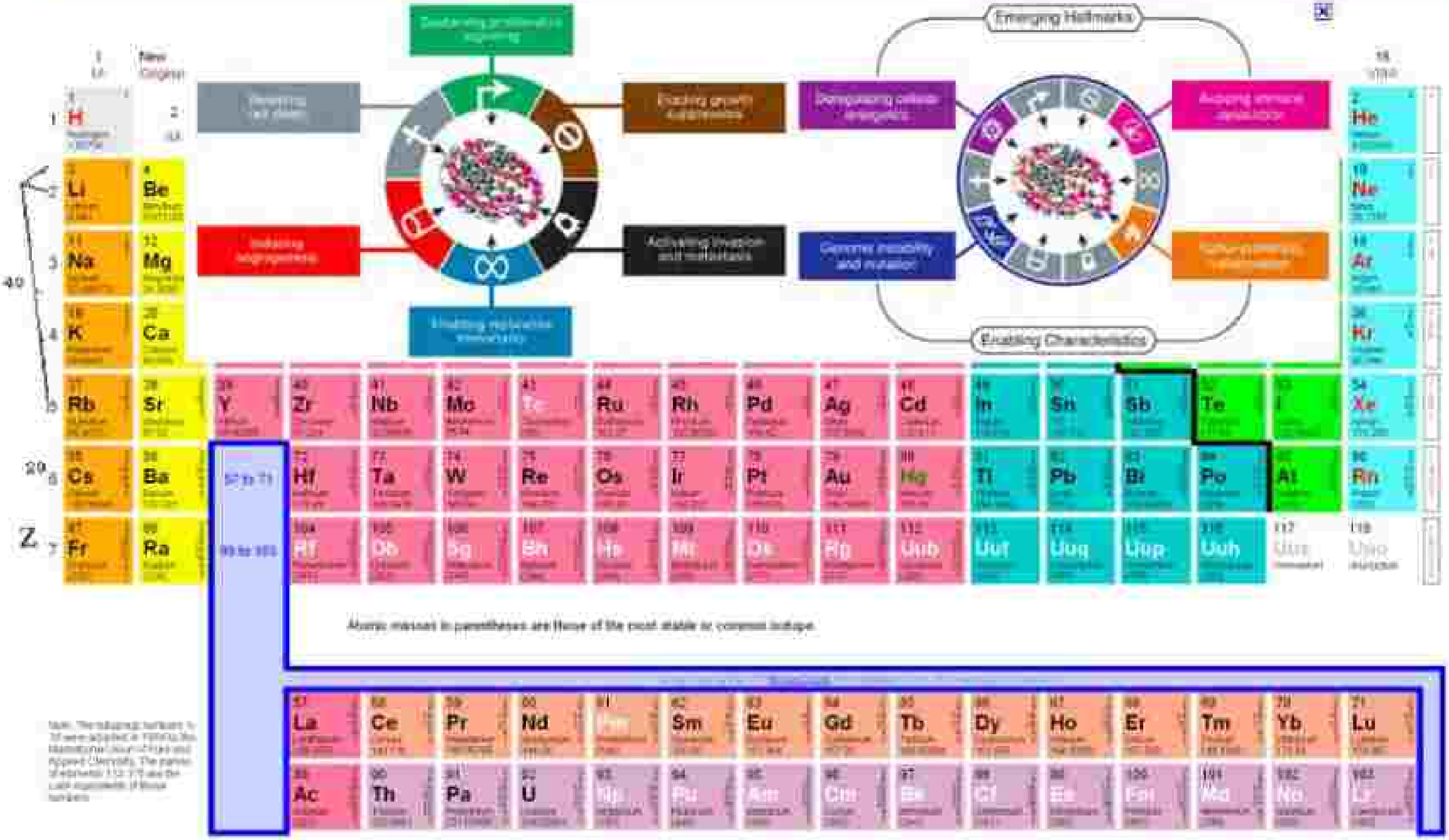
Genes mediating apoptosis, such as caspases, are identified as key drivers and enriched in the immune response module

33 key drivers in the [illegible] [illegible] interacting with [illegible] pathway, which directly interact with Par3 and aPKC

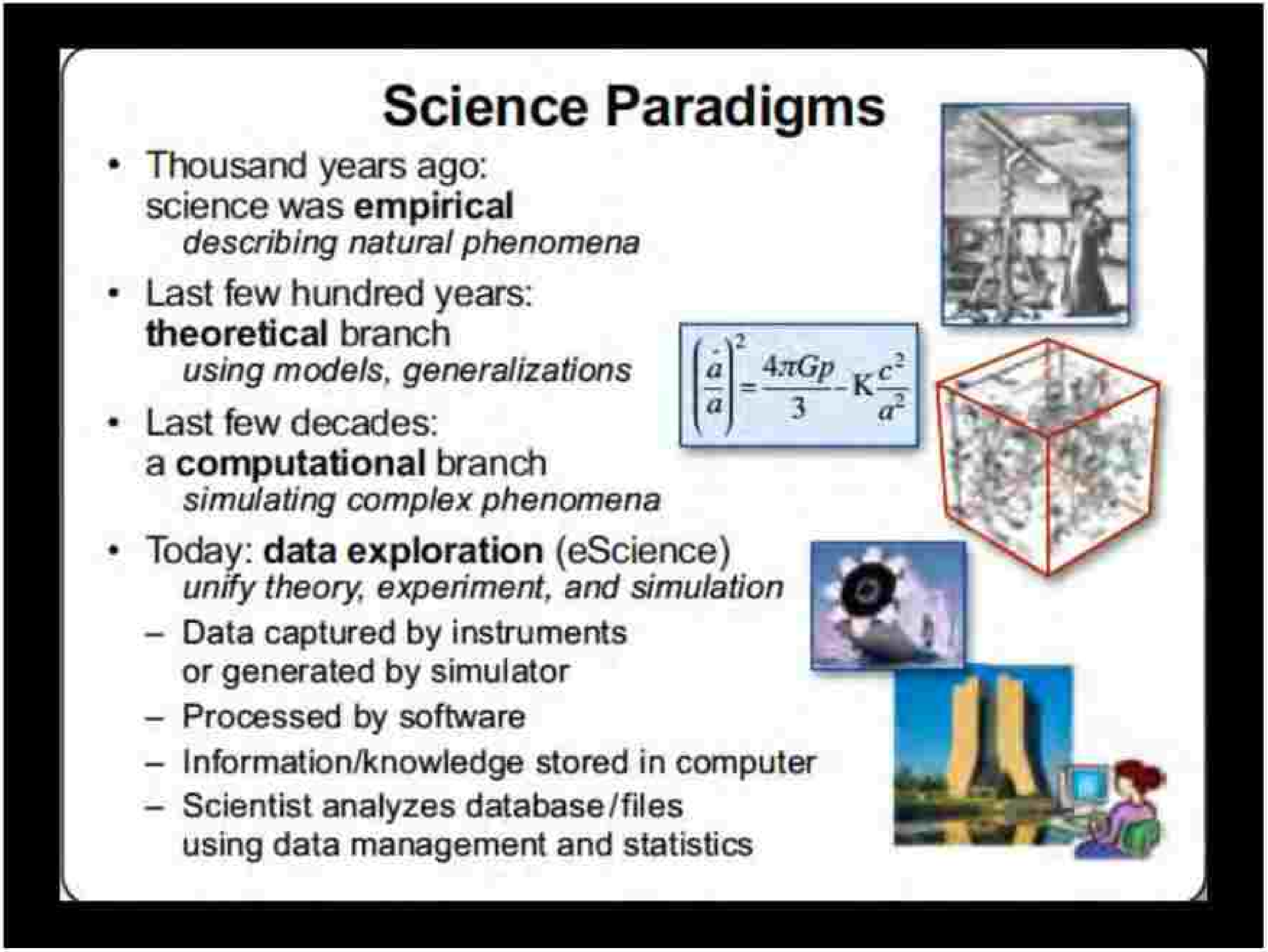
Key drivers identified in the cell cycle module, such as CDK4, Rb and p27 etc. are downstream of Ras/Raf/MAPK pathway

# The value of appropriate representations/ maps

## Periodic Table of the Elements



Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality
Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting inflammation
Enabling Characteristics
1
2
3
4
5
6
7
40
20
Z
H
He
Li
Be
Ne
Na
Mg
Ar
K
Ca
Kr
Rb
Sr
Y
Zr
Nb
Mo
Tc
Ru
Rh
Pd
Ag
Cd
In
Sn
Sb
Te
I
Xe
Cs
Ba
57 to 71
Hf
Ta
W
Re
Os
Ir
Pt
Au
Hg
Tl
Pb
Bi
Po
At
Rn
Fr
Ra
89 to 103
Rf
Db
Sg
Bh
Hs
Mt
Ds
Rg
Uub
Uut
Uuq
Uup
Uuh
Uus
Uuo
Atomic masses in parentheses are those of the most stable or common isotope.
La
Ce
Pr
Nd
Pm
Sm
Eu
Gd
Tb
Dy
Ho
Er
Tm
Yb
Lu
Ac
Th
Pa
U
Np
Pu
Am
Cm
Bk
Cf
Es
Fm
Md
No
Lr



# Science Paradigms

- Thousand years ago:
  science was **empirical**
  *describing natural phenomena*
- Last few hundred years:
  **theoretical** branch
  *using models, generalizations*
- Last few decades:
  a **computational** branch
  *simulating complex phenomena*
- Today: **data exploration** (eScience)
  *unify theory, experiment, and simulation*
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

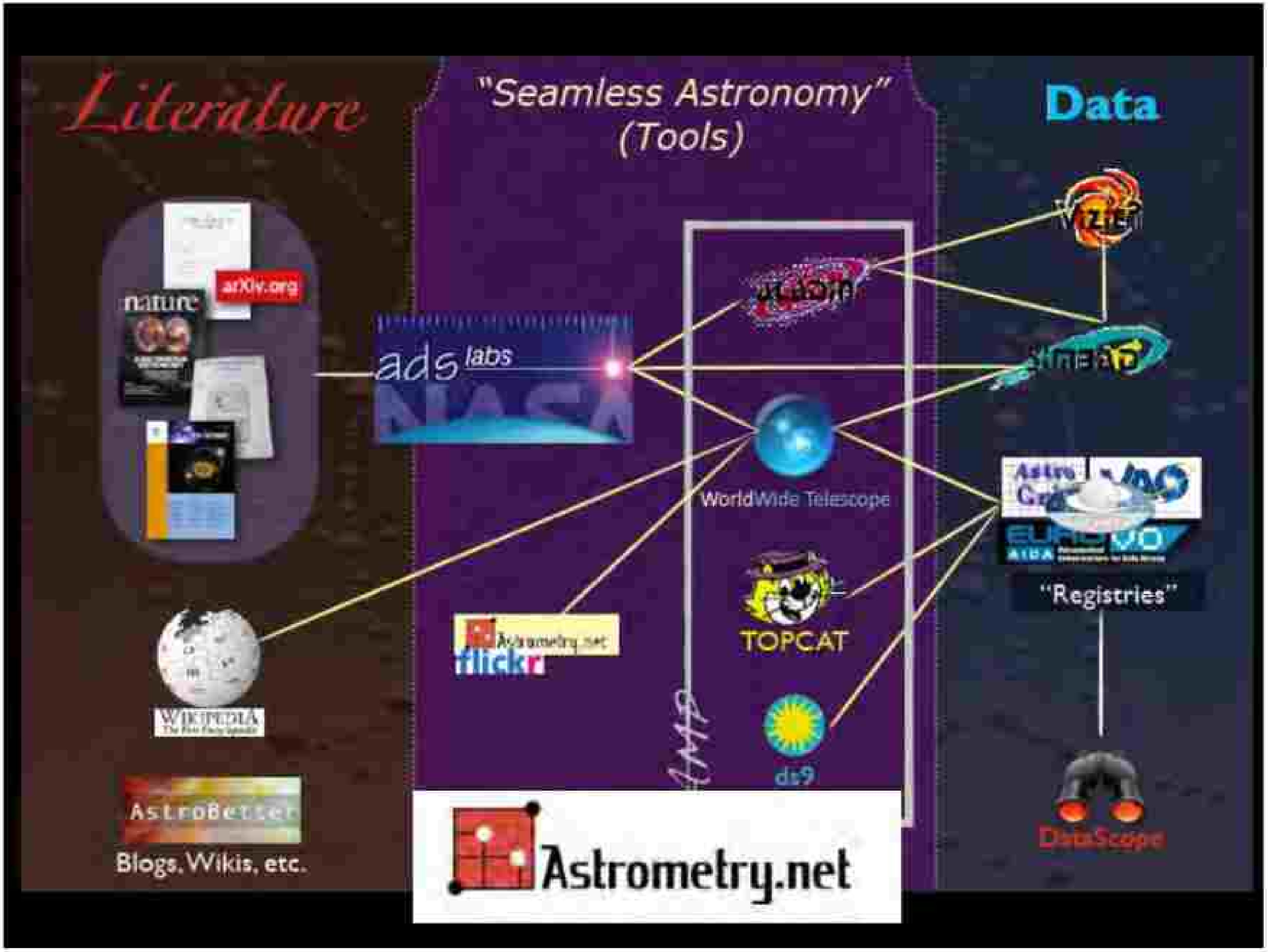
# “Data Intensive” Science- Fourth Scientific Paradigm

Equipment capable of generating massive amounts of data

IT Interoperability

Open Information System

Host evolving Models in a Compute Space- Knowledge Expert



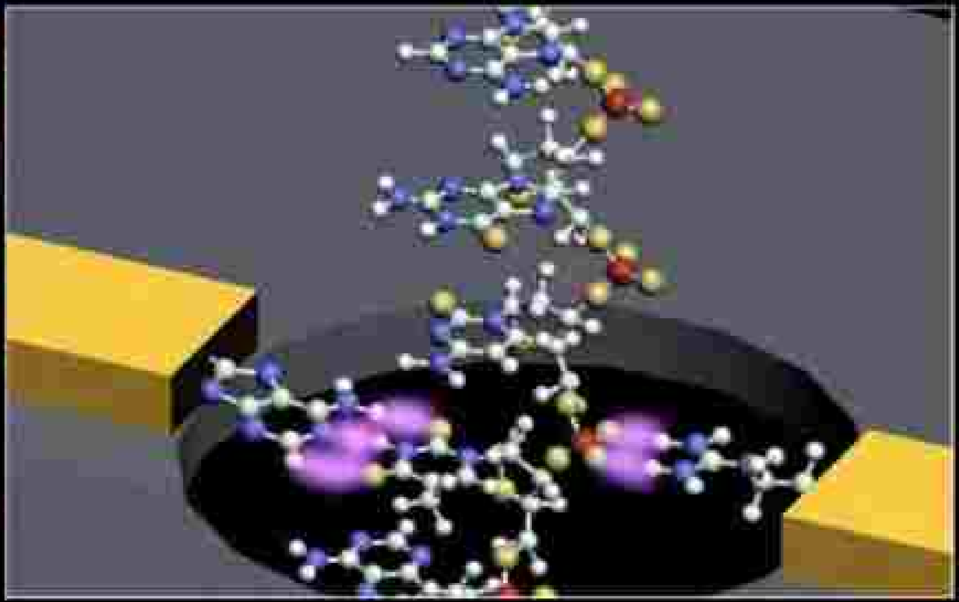
Literature

“Seamless Astronomy”
(Tools)

Data

# WHY NOT USE
“DATA INTENSIVE” SCIENCE
TO BUILD BETTER DISEASE MAPS?

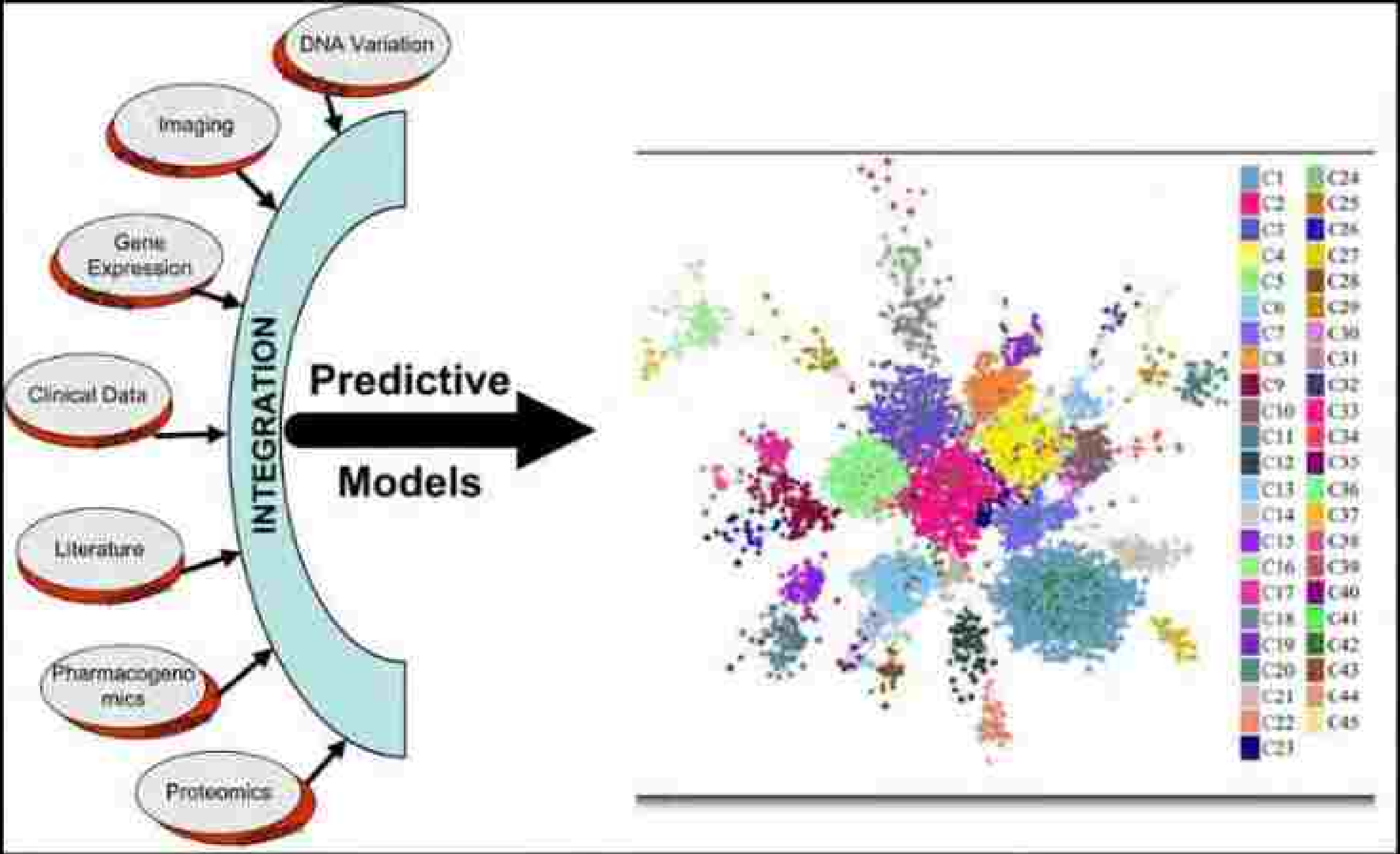
# what will it take to understand disease?



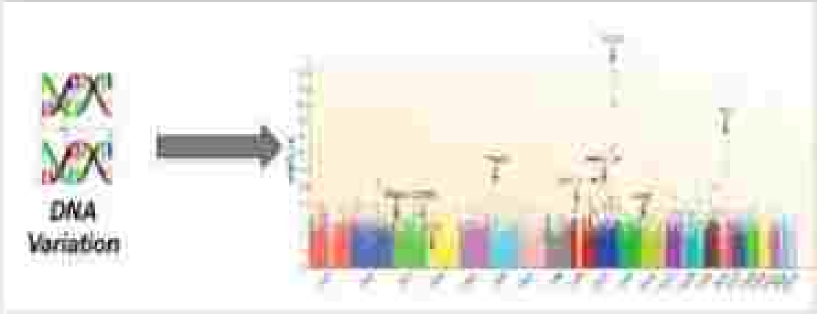
DNA  RNA PROTEIN (dark matter)

MOVING BEYOND ALTERED COMPONENT LISTS

# 2002 Can one build a “causal” model?

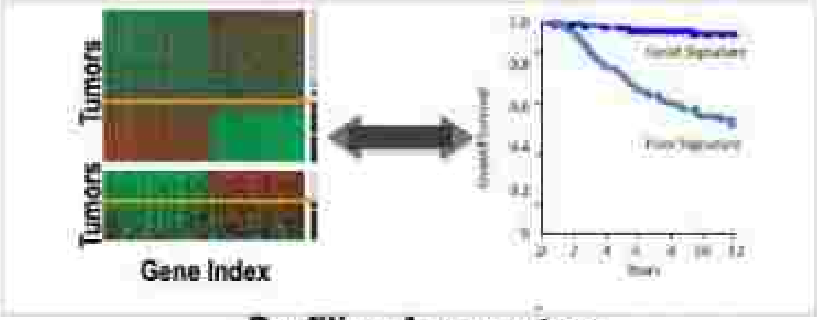


# How is genomic data used to understand biology?



Standard GWAS Approaches

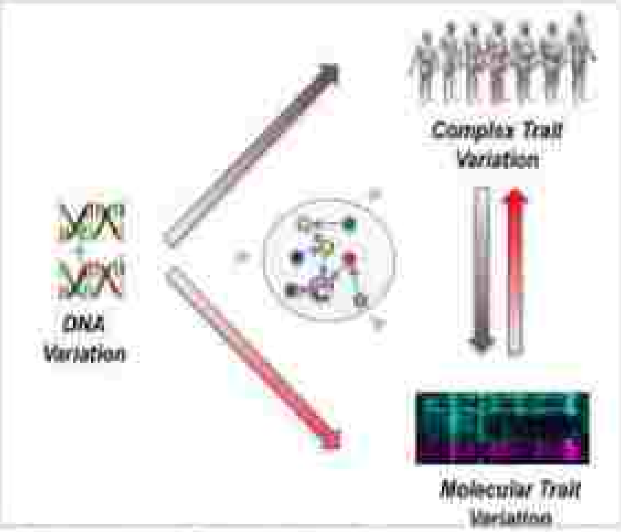
Identifies Causative DNA Variation but provides NO mechanism



Profiling Approaches

Genome scale profiling provide correlates of disease

➢ Many examples BUT what is cause and effect?

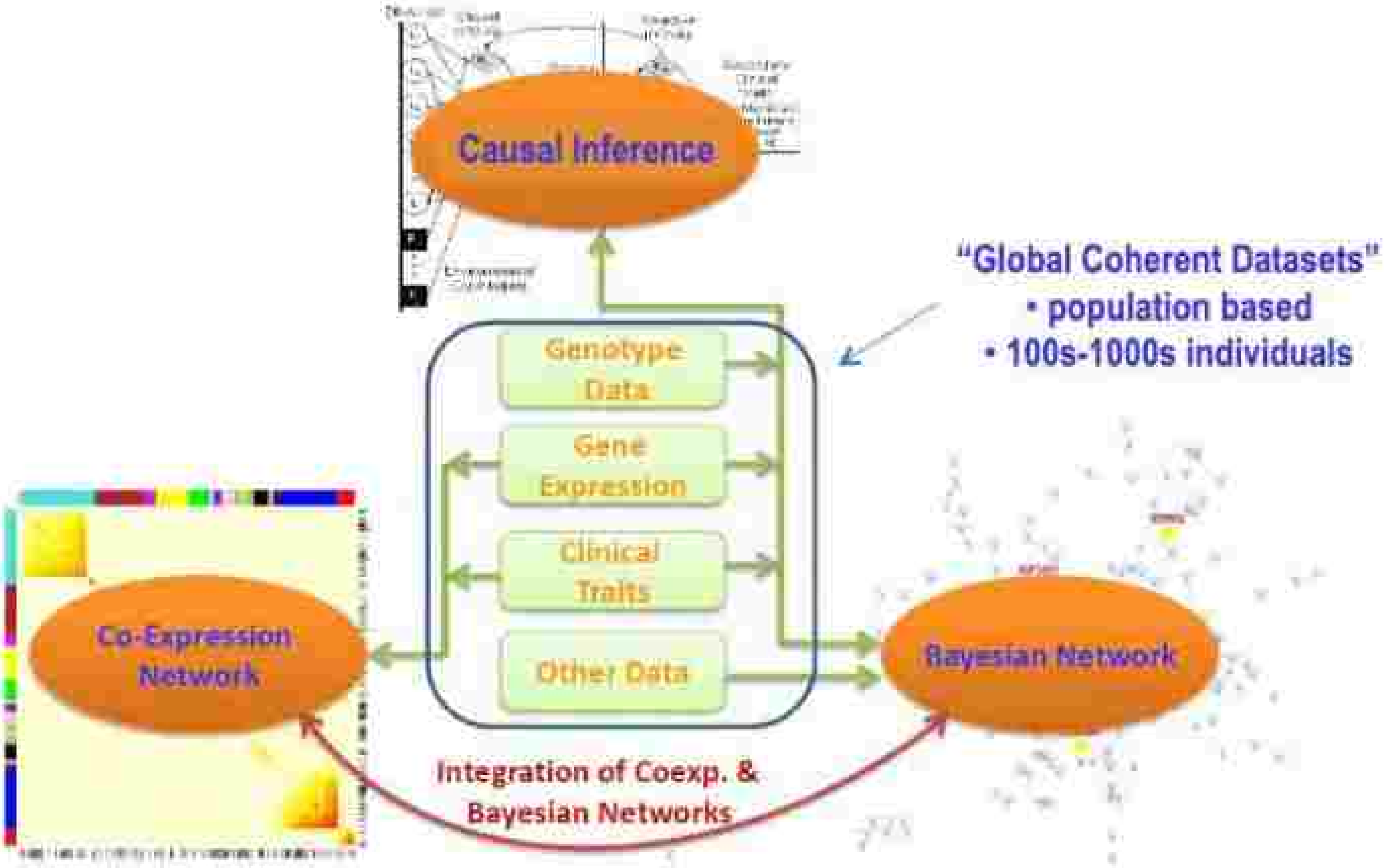


Integrated Genetics Approaches

- ➢ Provide unbiased view of molecular physiology as it relates to disease phenotypes
- ➢ Insights on mechanism
- ➢ Provide causal relationships and allows predictions

# Integration of Genotypic, Gene Expression & Trait Data

Schadt et al. Nature Genetics 37: 710 (2005)
Millstein et al. BMC Genetics 10: 23 (2009)



Causal Inference

Genotype Data

Gene Expression

Clinical Traits

Other Data

"Global Coherent Datasets"
- population based
- 100s-1000s individuals

Co-Expression Network

Bayesian Network

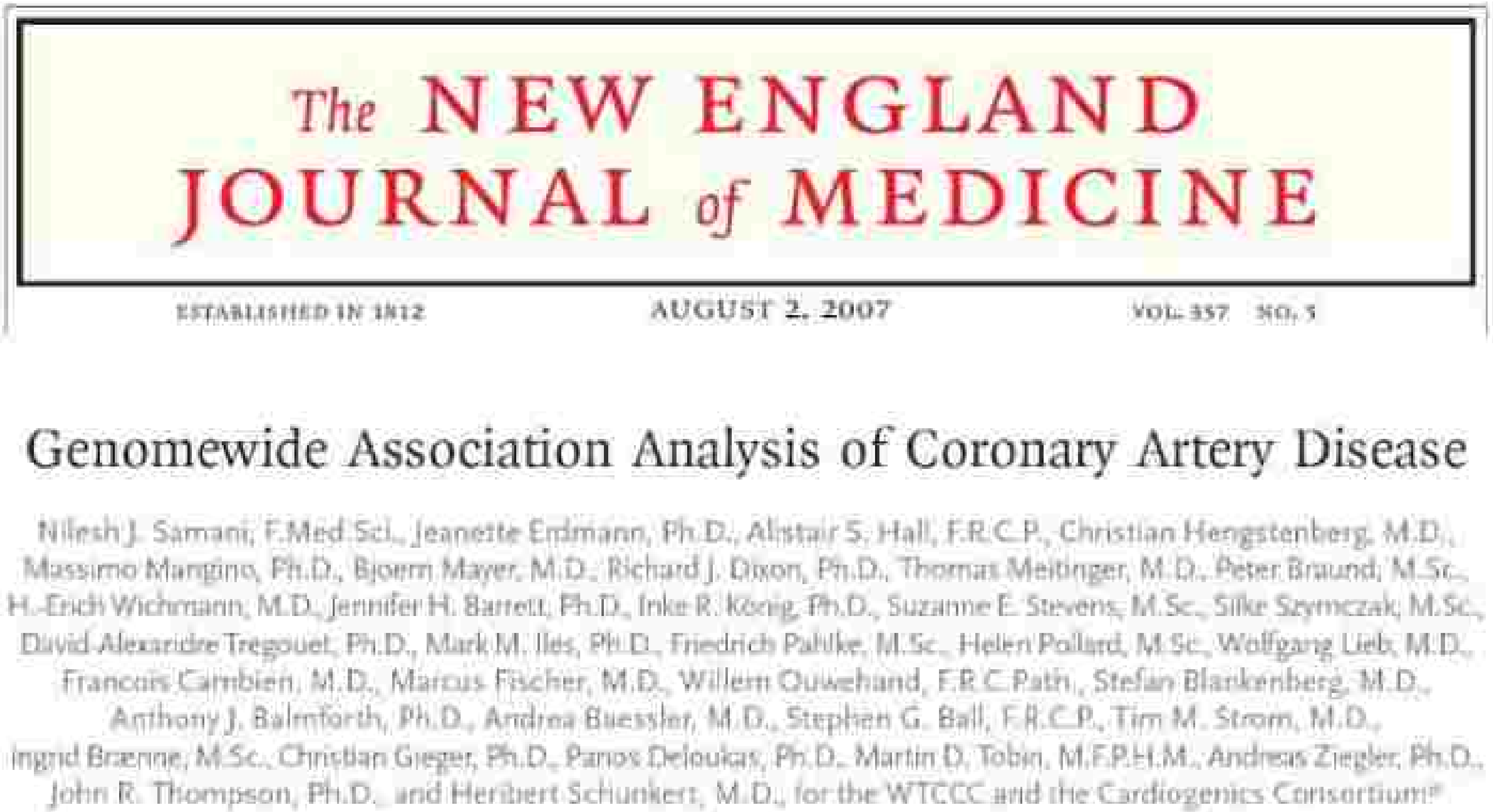
Integration of Coexp. & Bayesian Networks

Chen et al. Nature 452:429 (2008)
Zhang & Horvath. Stat.Appl.Genet.Mol.Biol. 4: article 17 (2005)

Zhu et al. Cytogenet Genome Res. 105:363 (2004)
Zhu et al. PLoS Comput. Biol. 3: e69 (2007)

# Association of SNPs at 1p13.3 with Coronary Artery Disease

## SNP rs599839 in the 1p13.3 locus associated with CAD: PSRC1 highlighted as candidate susceptibility gene

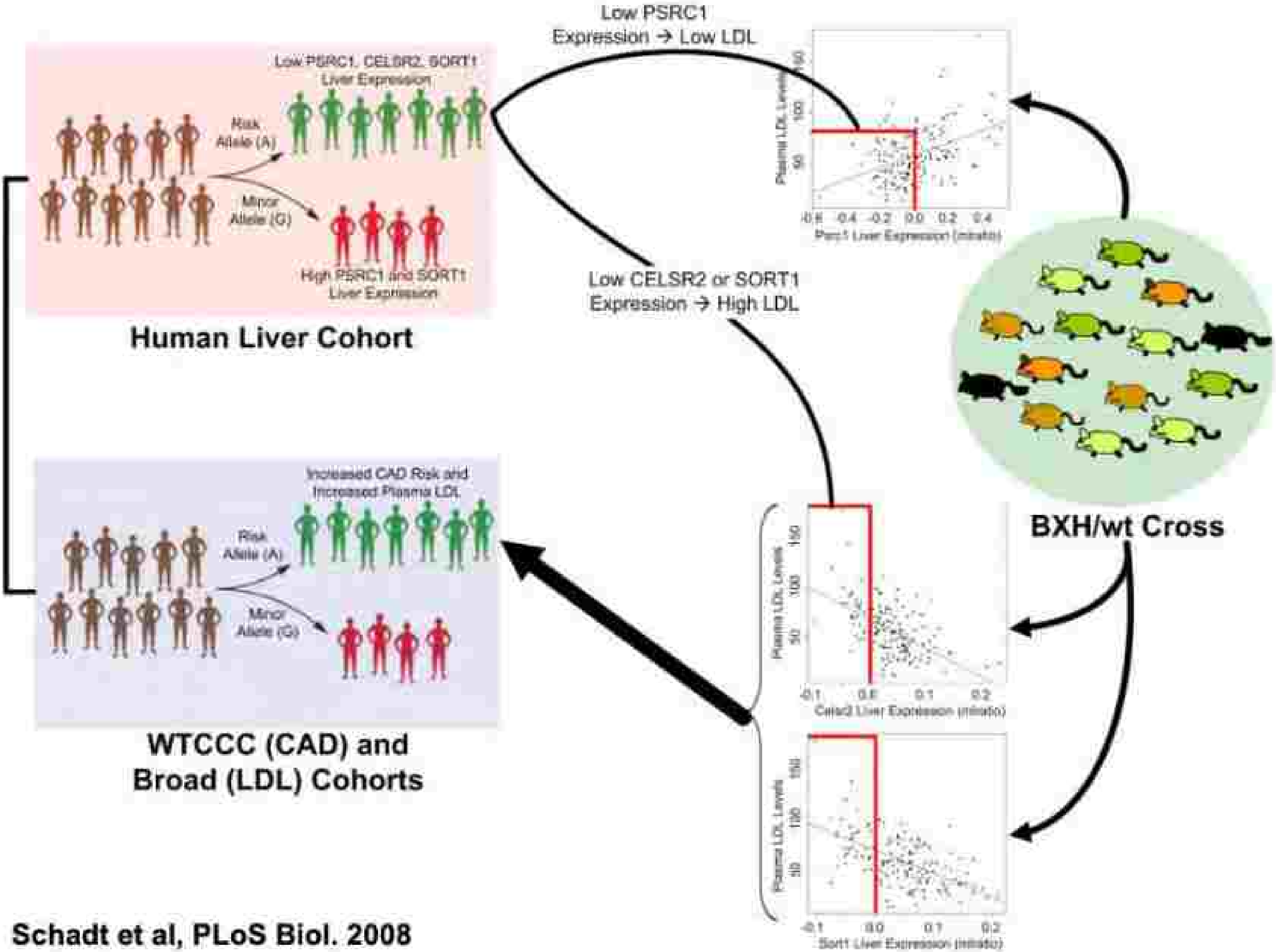


The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 2, 2007 VOL. 357 NO. 5

### Genomewide Association Analysis of Coronary Artery Disease

Nilesh J. Samani, F.Med.Sci., Jeanette Erdmann, Ph.D., Alistair S. Hall, F.R.C.P., Christian Hengstenberg, M.D., Massimo Mangino, Ph.D., Bjoern Mayer, M.D., Richard J. Dixon, Ph.D., Thomas Meitinger, M.D., Peter Braund, M.Sc., H.-Erich Wichmann, M.D., Jennifer H. Barrett, Ph.D., Inke R. König, Ph.D., Suzanne E. Stevens, M.Sc., Silke Szymczak, M.Sc., David-Alexandre Tregouet, Ph.D., Mark M. Iles, Ph.D., Friedrich Pahlke, M.Sc., Helen Pollard, M.Sc., Wolfgang Lieb, M.D., Francois Cambien, M.D., Marcus Fischer, M.D., Willem Ouwehand, F.R.C.Path., Stefan Blankenberg, M.D., Anthony J. Balmforth, Ph.D., Andrea Baessler, M.D., Stephen G. Ball, F.R.C.P., Tim M. Strom, M.D., Ingrid Brænne, M.Sc., Christian Gieger, Ph.D., Panos Deloukas, Ph.D., Martin D. Tobin, M.F.P.H.M., Andreas Ziegler, Ph.D., John R. Thompson, Ph.D., and Heribert Schunkert, M.D., for the WTCCC and the Cardiogenics Consortium*



Low PSRC1 Expression → Low LDL

Low CELSR2 or SORT1 Expression → High LDL

Human Liver Cohort

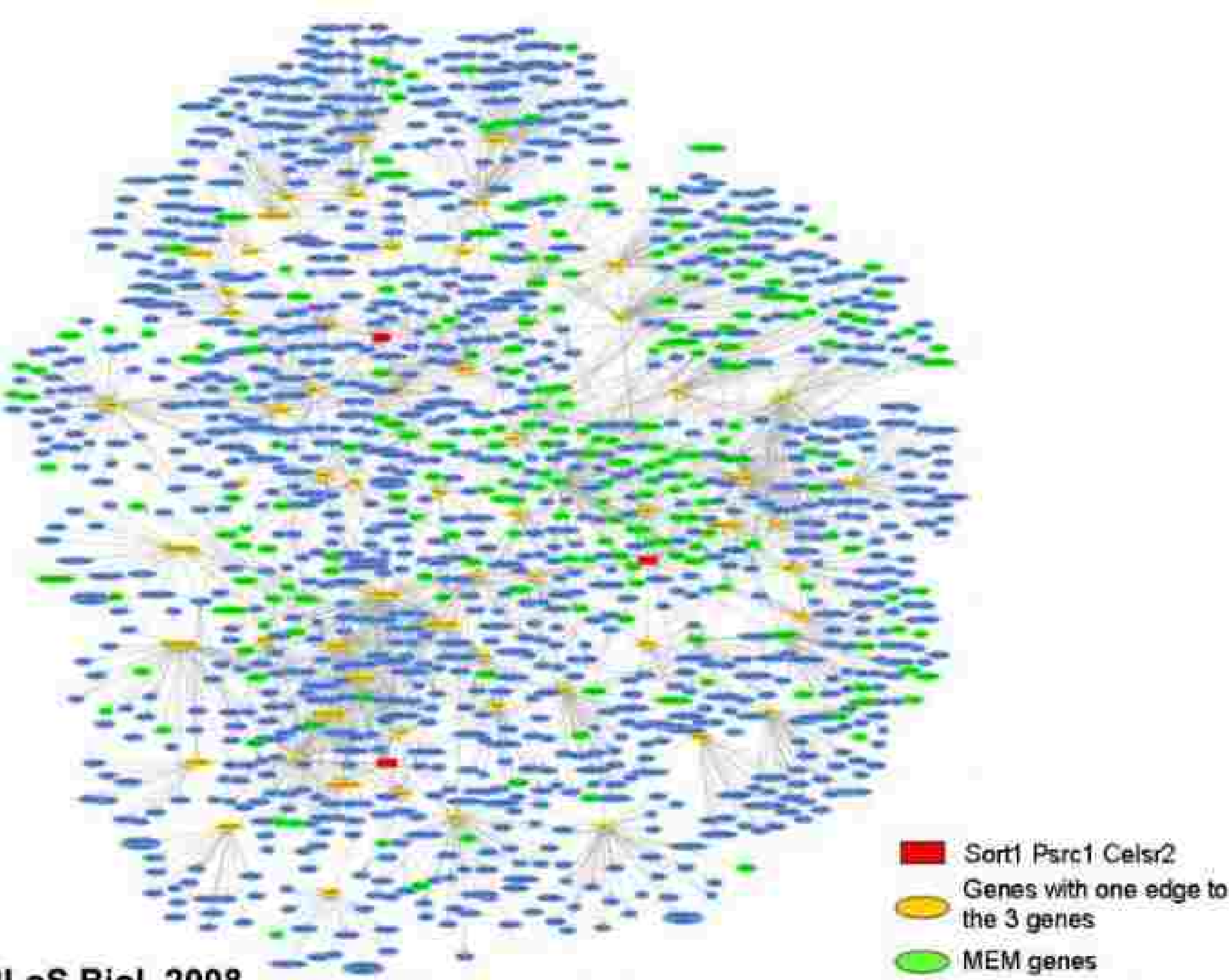
WTCCC (CAD) and Broad (LDL) Cohorts

BXH/wt Cross

Low PSRC1, CELSR2, SORT1 Liver Expression
Risk Allele (A)
Minor Allele (G)
High PSRC1 and SORT1 Liver Expression
Increased CAD Risk and Increased Plasma LDL
Risk Allele (A)
Minor Allele (G)
Plasma LDL Levels
50
100
150
-0.6
-0.4
-0.2
0.0
0.2
0.4
Psrc1 Liver Expression (mlratio)
Plasma LDL Levels
50
100
150
-0.1
0.0
0.1
0.2
Celsr2 Liver Expression (mlratio)
Plasma LDL Levels
50
100
150
-0.1
0.0
0.1
0.2
Sort1 Liver Expression (mlratio)

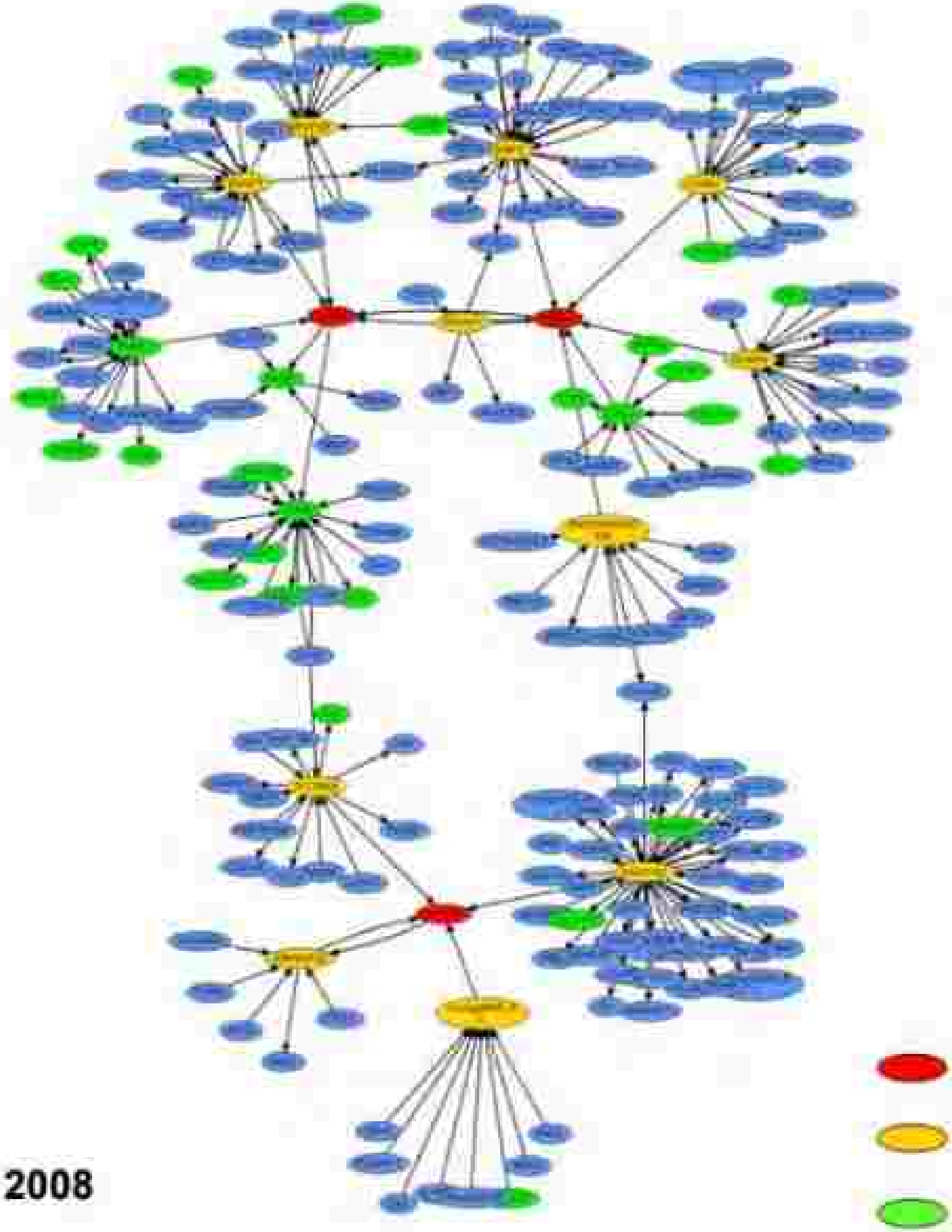
Schadt et al, PLoS Biol. 2008

# Mouse network around Sort1, Psrc1, and Celsr2



Schadt et al, PLoS Biol. 2008

# Human network around Sort1, Psrc1, and Celsr2



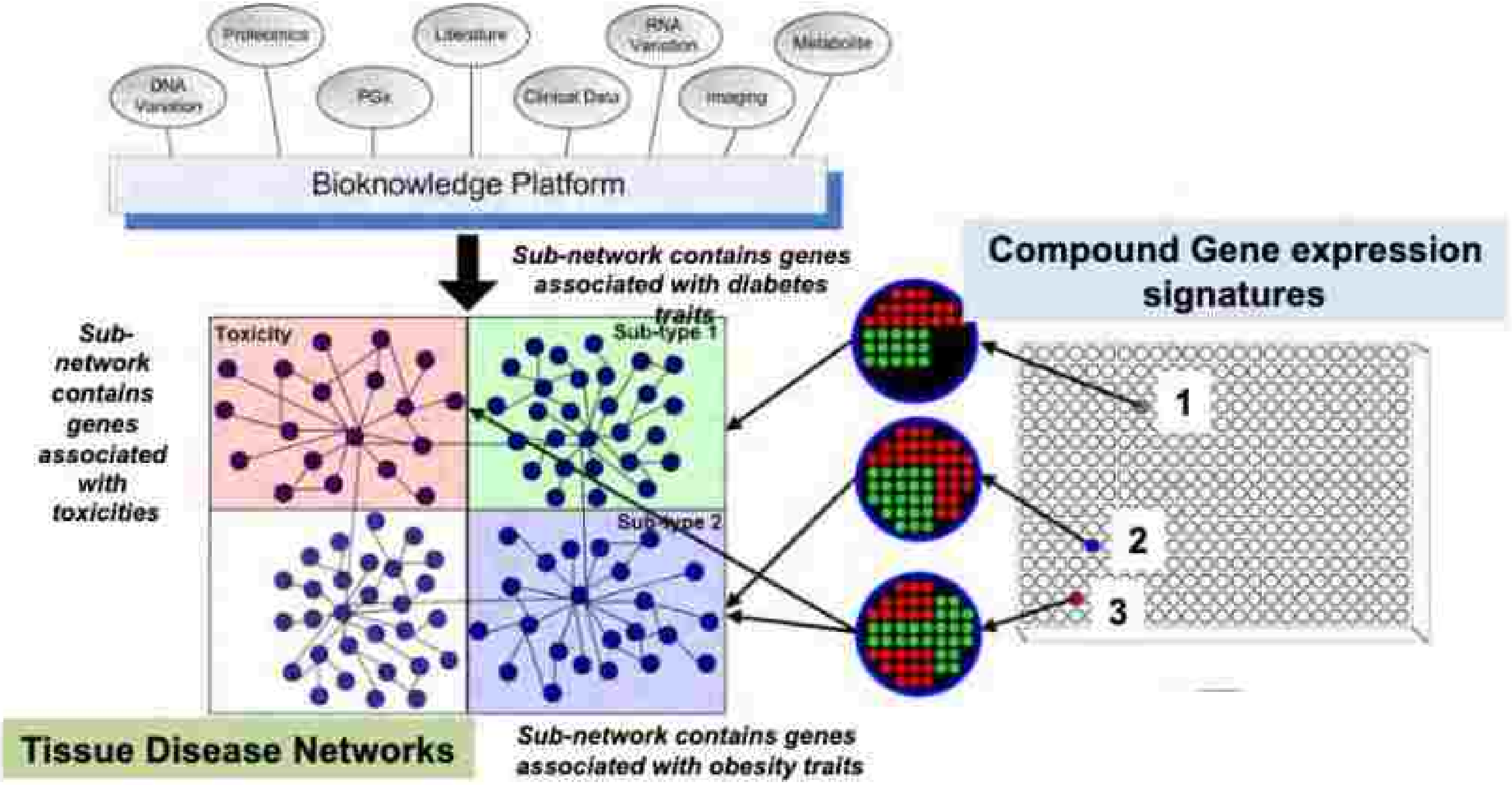
🟥 Sort1 Psrc1 Celsr2

🟨 Genes with one edge to the 3 genes

🟩 MEM genes

**Schadt et al, PLoS Biol. 2008**

# Map compound signatures to disease networks



Bioknowledge Platform

*Sub-network contains genes associated with diabetes traits*

Compound Gene expression signatures

*Sub-network contains genes associated with toxicities*

Tissue Disease Networks

*Sub-network contains genes associated with obesity traits*

Compound 1: Drug signature significantly enriched in subnetwork associated with diabetes traits

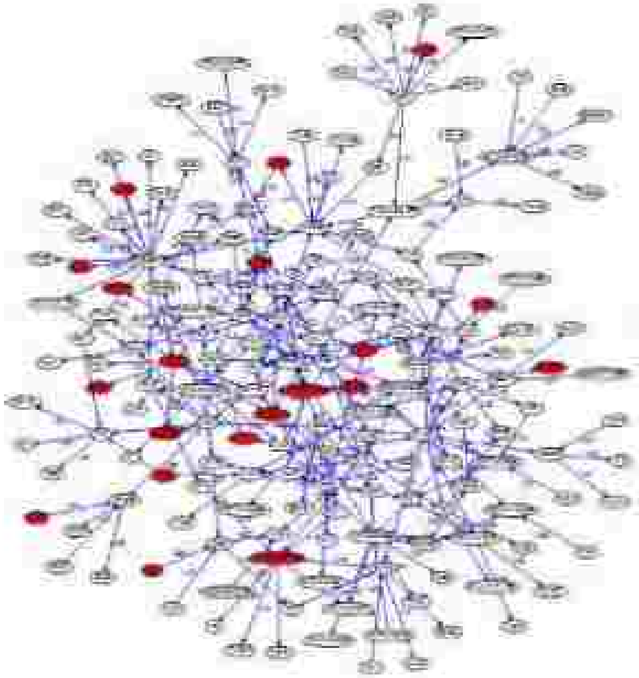
Compound 2: Drug signature significantly enriched in subnetwork associated with obesity traits

Compound 3: Drug signature significantly enriched in subnetwork associated with obesity traits BUT also in subnetwork associated with toxicities

# Case Study – Target A/Drug B

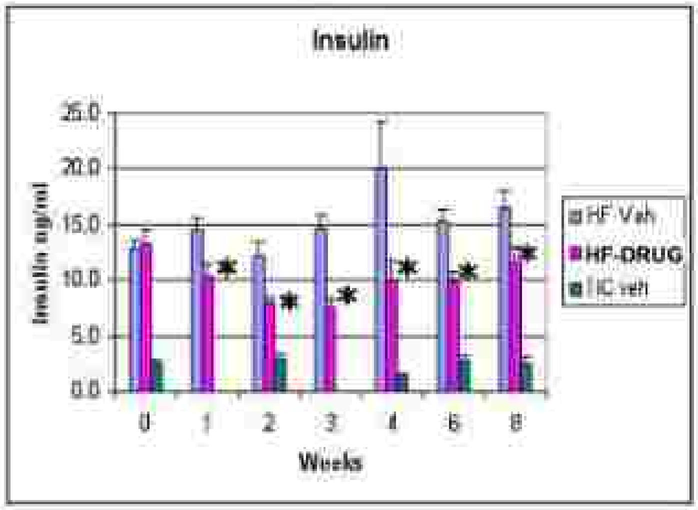
**NO CELL DYNAMICS NEEDED**

**Identified compound whose signature significantly intersected with Islet module**

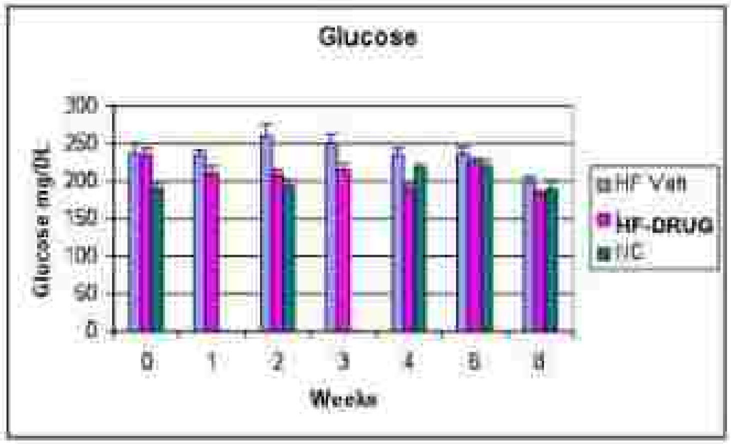


- **Test carried out in a Diet-Induced Obesity model on the B6 background**
  - **Model for obesity and insulin resistance**
- **Animals treated with compound over an 8 week interval, starting at 8 weeks of age**
- ***No significant Adverse Events in 30 day human clinical trial for another indication***

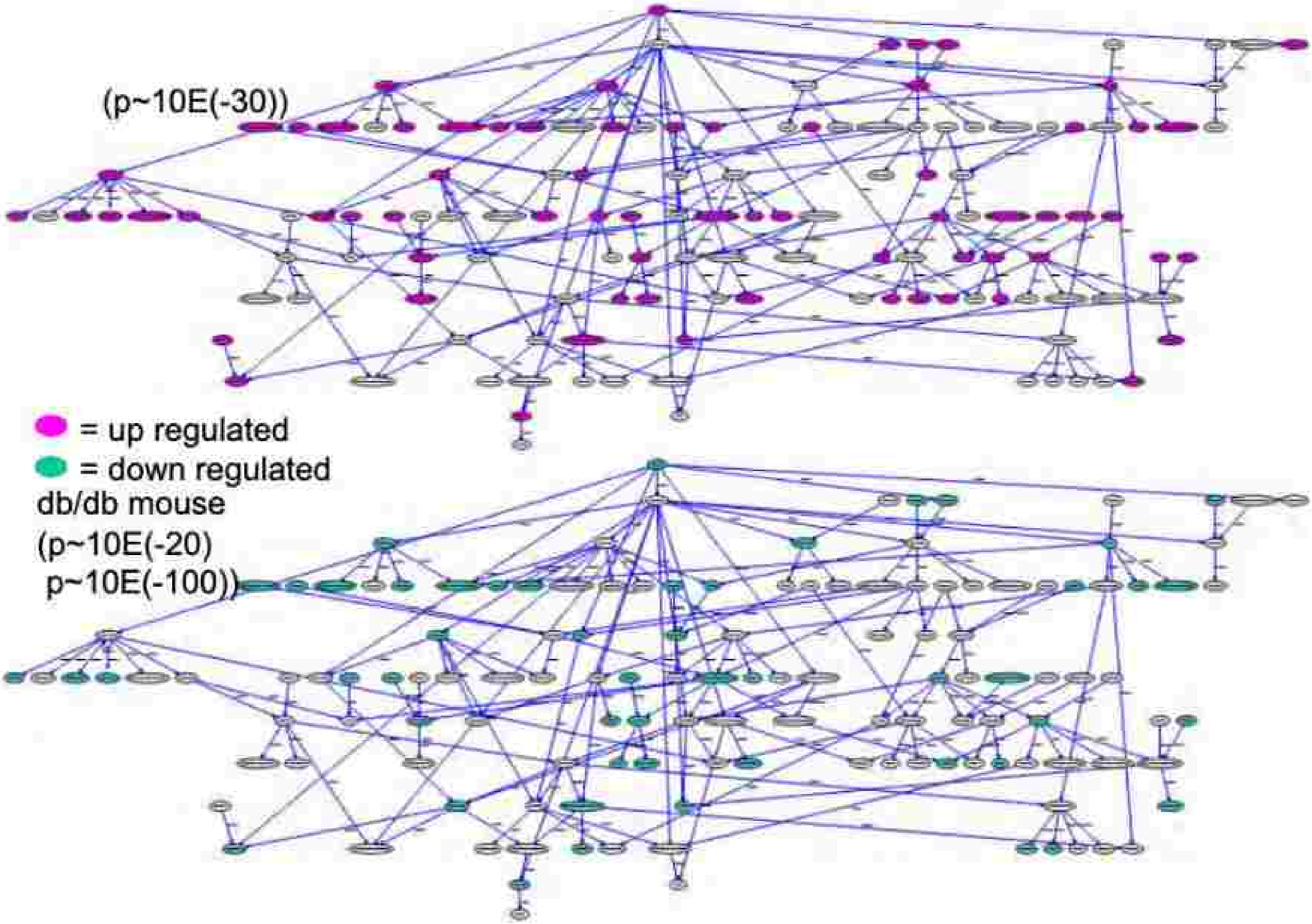
**Fasting Insulin**



**Fasting Glucose**



# Our ability to integrate compound data into our network analyses



(p~10E(-30))

● = up regulated

● = down regulated

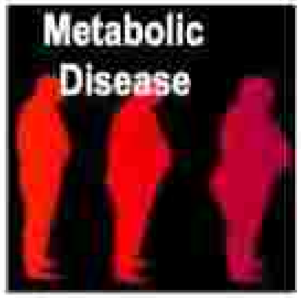
db/db mouse

(p~10E(-20)

p~10E(-100))

# Extensive Publications now Substantiating Scientific Approach
# Probabilistic Causal Bionetwork Models

- **>80 Publications from Rosetta Genetics Group (~30 scientists) over 5 years including high profile papers in PLoS Nature and Nature Genetics**



*"Genetics of gene expression surveyed in maize, mouse and man."* **Nature**. (2003)

*"Variations in DNA elucidate molecular networks that cause disease."* **Nature**. (2008)

*"Genetics of gene expression and its effect on disease."* **Nature**. (2008)

*"Validation of candidate causal genes for obesity that affect..."* **Nat Genet**. (2009)

**..... Plus 10 additional papers in Genome Research, PLoS Genetics, PLoS Comp.Biology, etc**

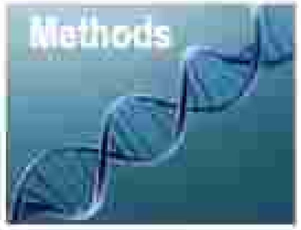
***"Identification of pathways for atherosclerosis."* Circ Res. (2007)**

*"Mapping the genetic architecture of gene expression in human liver."* **PLoS Biol**. (2008)

**...... Plus 5 additional papers in Genome Res., Genomics, Mamm.Genome**

*"Integrating genotypic and expression data ...for bone traits..."* **Nat Genet.** (2005)

*..approach to identify candidate genes regulating BMD..."* **J Bone Miner Res**. (2009)



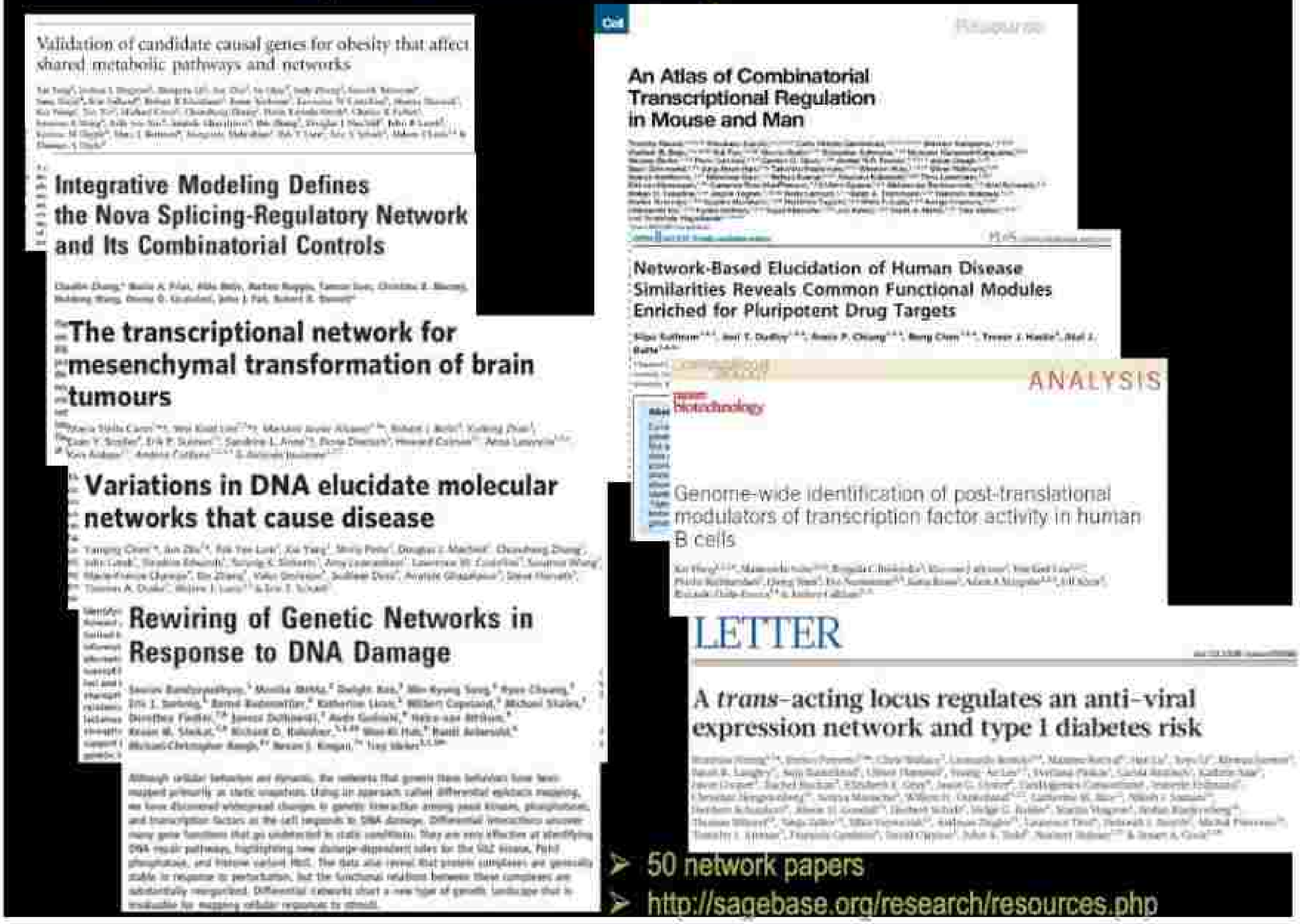
***"An integrative genomics approach to infer causal associations ...Nat Genet . (2005)***

***"Increasing the power to detect causal associations...* PLoS Comput Biol. (2007)**

*"Integrating large-scale functional genomic data ..."* **Nat Genet.** (2008)

...... Plus 3 additional papers in **PLoS Genet., BMC Genet.**

# List of Influential Papers in Network Modeling



Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks

[illegible]

Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls

Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, Robert B. Darnell

The transcriptional network for mesenchymal transformation of brain tumours

Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J. Bollo, Xudong Zhao, Evan Y. Snyder, Erik P. Sulman, Sandrine L. Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano & Antonio Iavarone

Variations in DNA elucidate molecular networks that cause disease

Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J. MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K. Sieberts, Amy Leonardson, Lawrence W. Castellini, Susanna Wang, Marie-France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A. Drake, Aldons J. Lusis & Eric E. Schadt

Rewiring of Genetic Networks in Response to DNA Damage

Sourav Bandyopadhyay, Monika Mehta, Dwight Kuo, Min-Kyung Sung, Ryan Chuang, Eric J. Jaehnig, Bernd Bodenmiller, Katherine Licon, Wilbert Copeland, Michael Shales, Dorothea Fiedler, Janusz Dutkowski, Aude Guénolé, Haico van Attikum, Kevan M. Shokat, Richard D. Kolodner, Won-Ki Huh, Ruedi Aebersold, Michael-Christopher Keogh, Nevan J. Krogan, Trey Ideker

Although cellular behaviors are dynamic, the networks that govern these behaviors have been mapped primarily as static snapshots. Using an approach called differential epistasis mapping, we have discovered widespread changes in genetic interaction among yeast kinases, phosphatases, and transcription factors as the cell responds to DNA damage. Differential interactions uncover many gene functions that go undetected in static conditions. They are very effective at identifying DNA repair pathways, highlighting new damage-dependent roles for the Slt2 kinase, Pph3 phosphatase, and histone variant Htz1. The data also reveal that protein complexes are generally stable in response to perturbation, but the functional relations between these complexes are substantially reorganized. Differential networks chart a new type of genetic landscape that is invaluable for mapping cellular responses to stimuli.

Cell

Resource

An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man

[illegible]

PLoS

Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets

Silpa Suthram, Joel T. Dudley, Annie P. Chiang, Rong Chen, Trevor J. Hastie, Atul J. Butte

nature biotechnology

ANALYSIS

Genome-wide identification of post-translational modulators of transcription factor activity in human B cells

Kai Wang, Masumichi Saito, Bogdan C. Bisikirska, Mariano J. Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A. Margolin, Ulf Klein, Riccardo Dalla-Favera & Andrea Califano

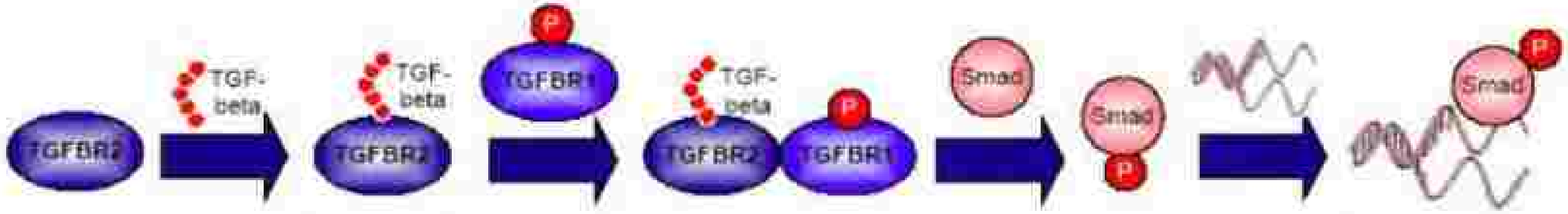
LETTER

A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk

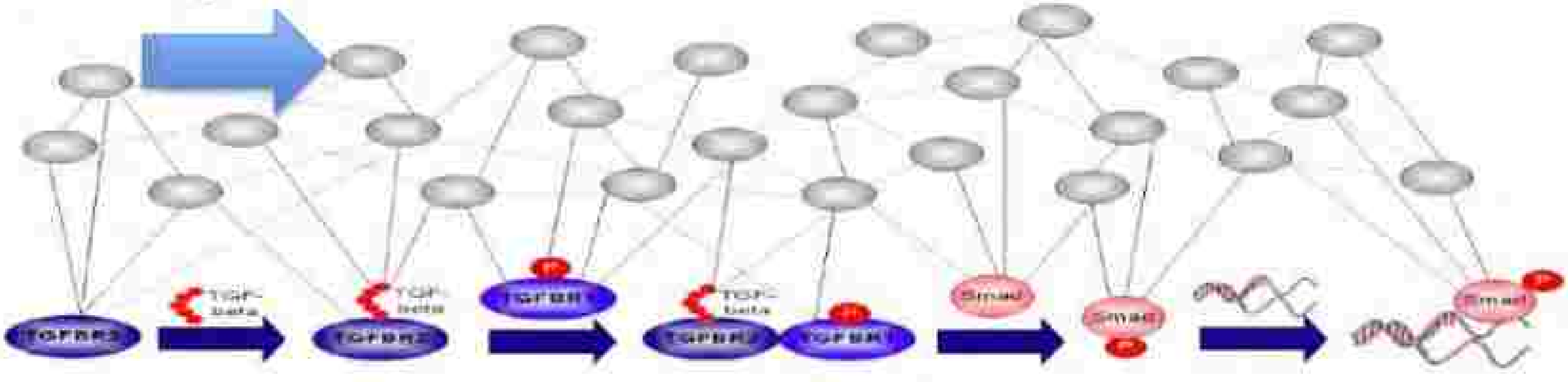
[illegible]

- 50 network papers
- http://sagebase.org/research/resources.php

## *The way we like to think*:



## *The way it is:*



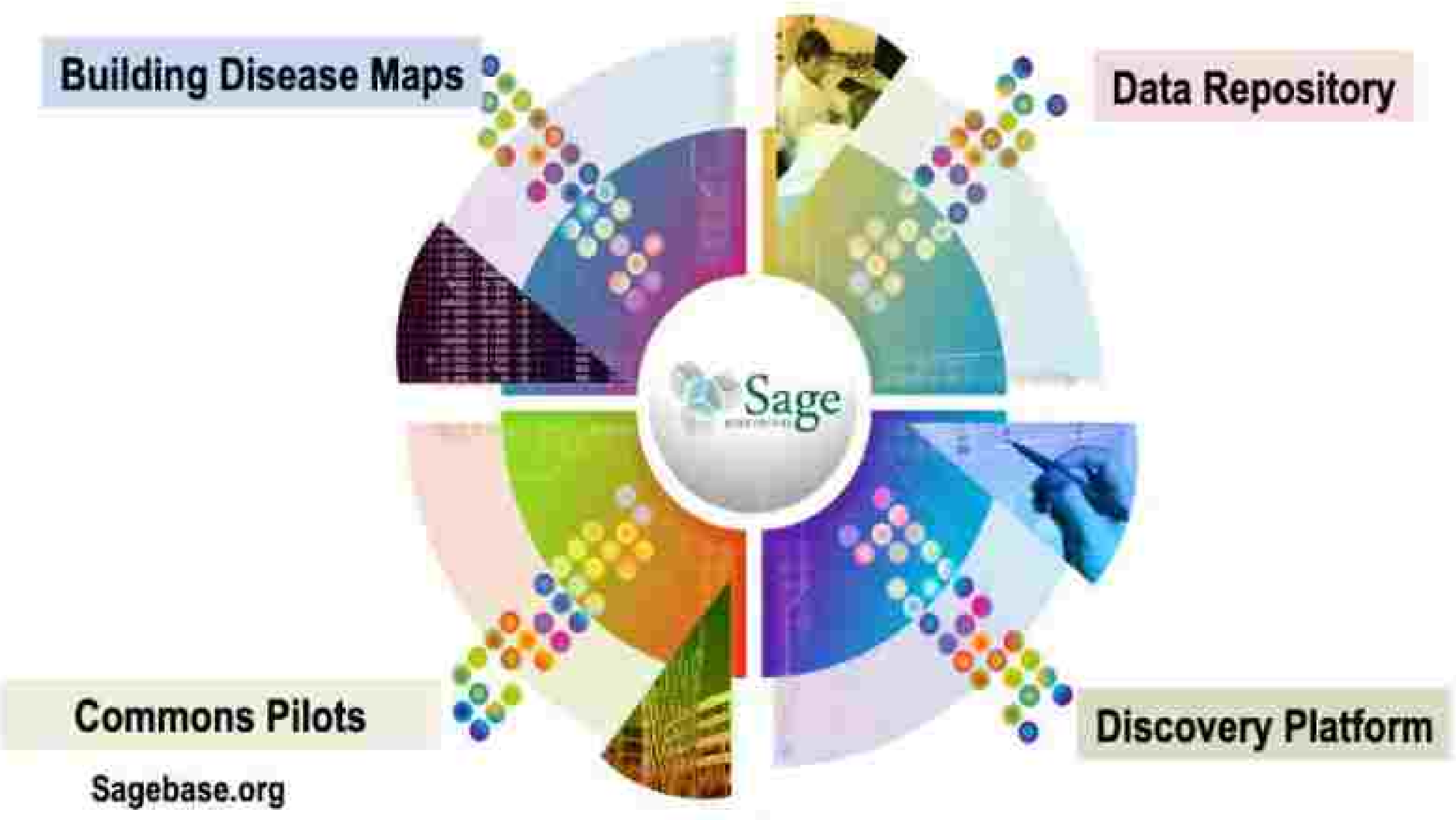
**(Eric Schadt)**

Recognition that the benefits of bionetwork based molecular models of diseases are powerful but that they **require significant resources**

Appreciation that it will **require decades** of evolving representations as real complexity emerges and needs to be integrated with therapeutic interventions

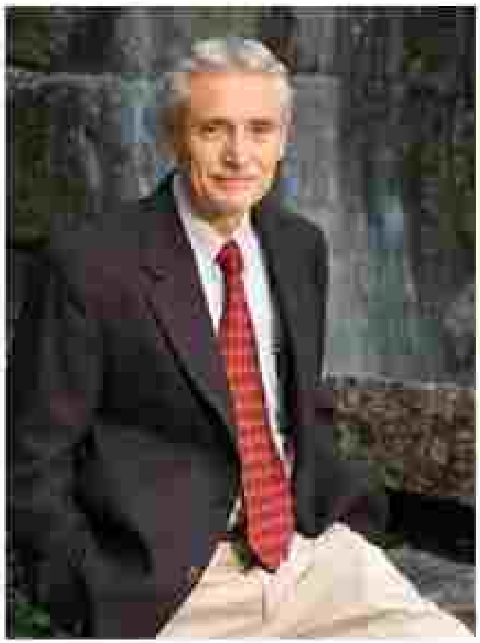
# Sage Mission

Sage Bionetworks is a non-profit organization with a vision to create a commons where integrative bionetworks are evolved by contributor scientists with a shared vision to accelerate the elimination of human disease



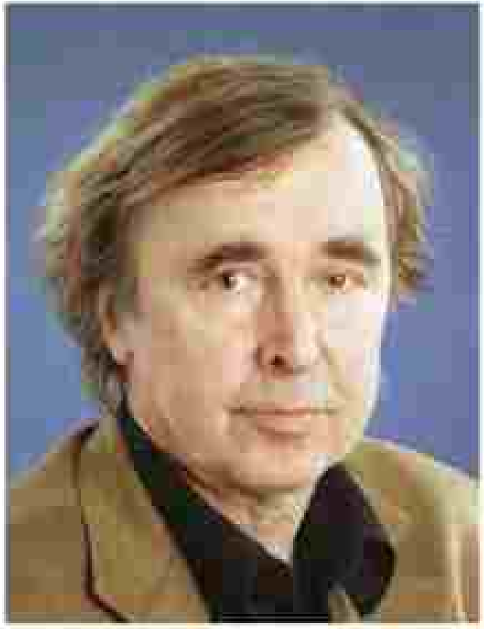
Sagebase.org

# Board of Directors- Sage Bionetworks



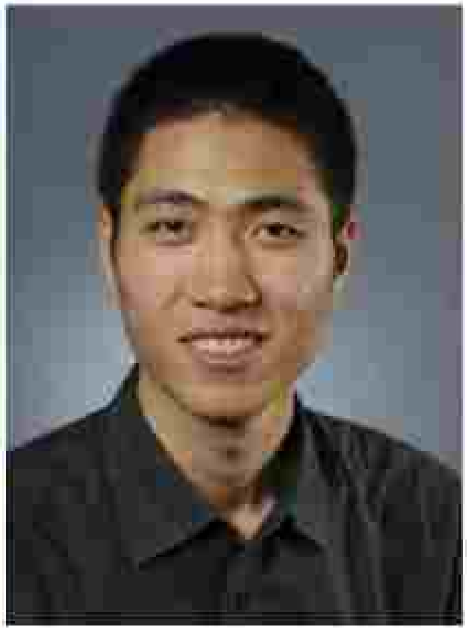
**Lee Hartwell**

**Ex President FHCRC**
**Co-Founder Rosetta**



**Hans Wizgell**

**ExPresident Karolinska**
**Head SAB Rosetta**



**WangJun**

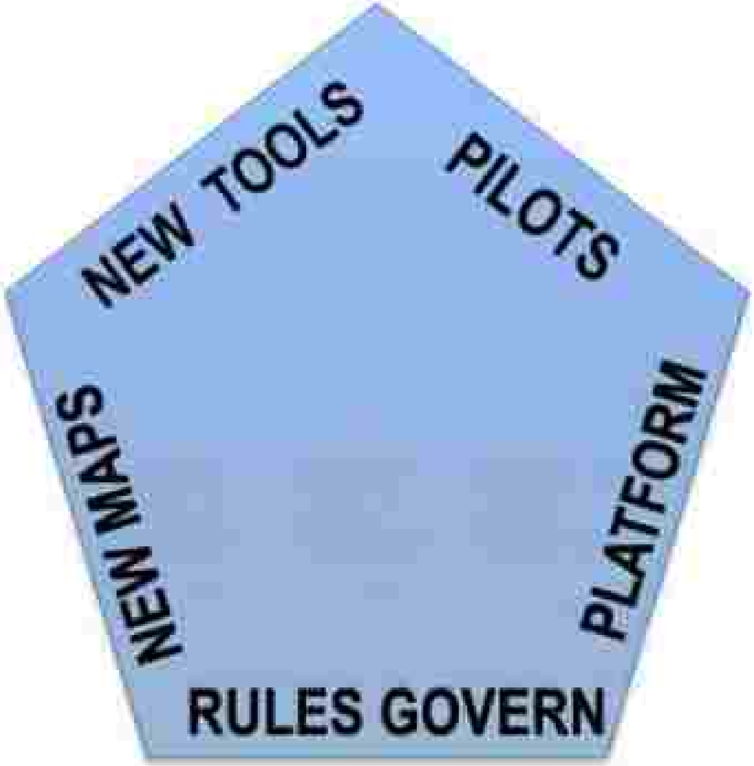
**Executive Director**
**BGI**

**Jeff Hammerbacher**

**CEO Cloudera**
**Built and Headed**
**Facebook**
**Data Architecture**

# Sage Bionetworks Collaborators

- Pharma Partners
  - Merck, Pfizer, Takeda, Astra Zeneca, Amgen, Johnson &Johnson
- Foundations
  - Kauffman CHDI, Gates Foundation
- Government
  - NIH, LSDF
- Academic
  - Levy (Framingham)
  - Rosengren (Lund)
  - Krauss (CHORI)
- Federation
  - Ideker, Califarno, Butte, Schadt



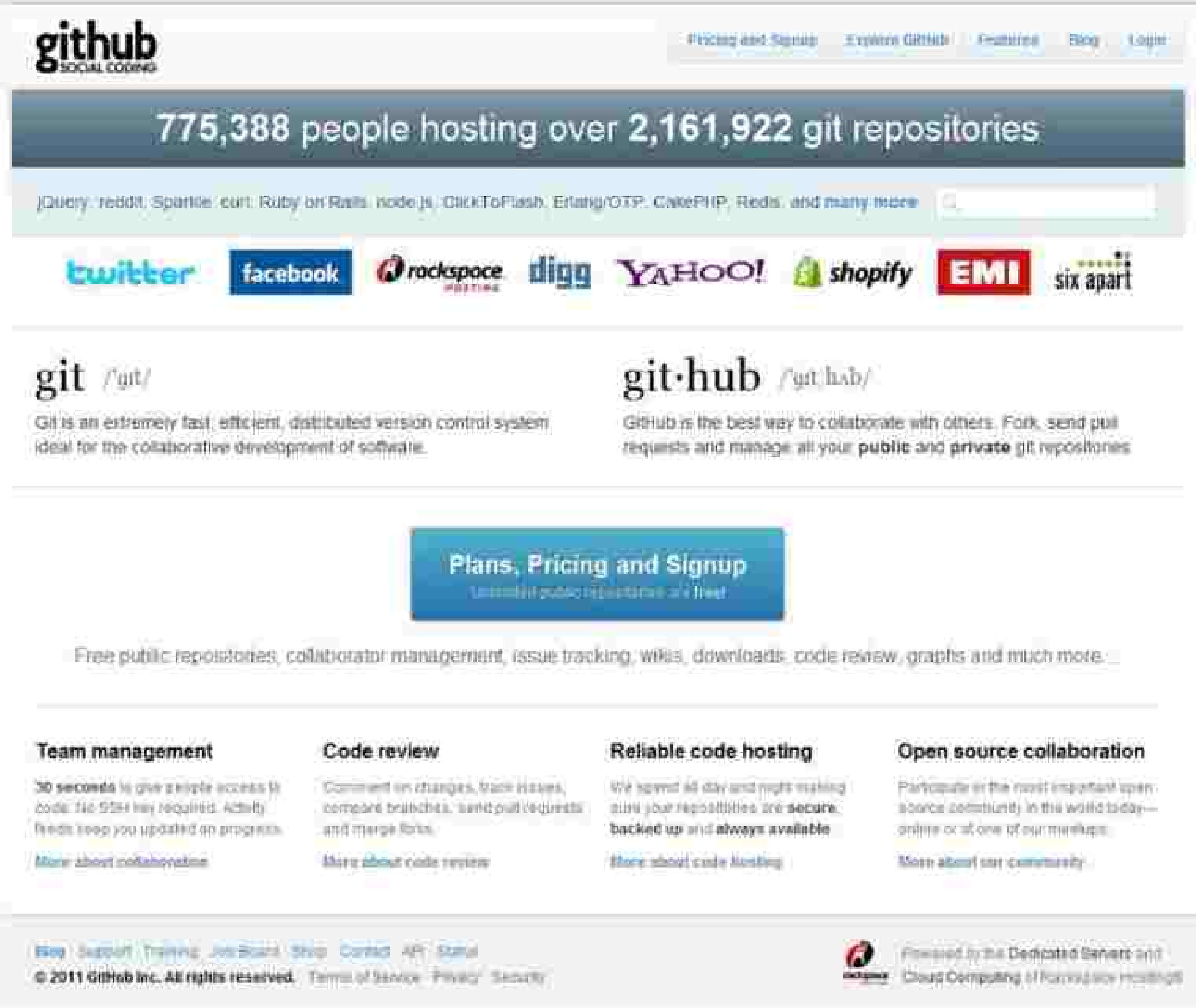
PLATFORM
Sage Platform and Infrastructure Builders-
( Academic Biotech and Industry IT Partners...)

PILOTS= PROJECTS FOR COMMONS
Data Sharing Commons Pilots-
(Federation, CCSB, Inspire2Live....)

NEW TOOLS
Data Tool and Disease Map Generators-
(Global coherent data sets, Cytoscape,
Clinical Trialists, Industrial Trialists, CROs…)

NEW MAPS
Disease Map and Tool Users-
( Scientists, Industry, Foundations, Regulators...)

RULES AND GOVERNANCE
Data Sharing Barrier Breakers-
(Patients Advocates, Governance
and Policy Makers, Funders...)



775,388 people hosting over 2,161,922 git repositories

jQuery, reddit, Sparkle, curl, Ruby on Rails, node.js, ClickToFlash, Erlang/OTP, CakePHP, Redis, and many more

git /gɪt/

Git is an extremely fast, efficient, distributed version control system ideal for the collaborative development of software.

git·hub /ˈgɪtˌhʌb/

GitHub is the best way to collaborate with others. Fork, send pull requests and manage all your **public** and **private** git repositories.

Plans, Pricing and Signup
Unlimited public repositories are free!

Free public repositories, collaborator management, issue tracking, wikis, downloads, code review, graphs and much more…

**Team management**

**30 seconds** to give people access to code. No SSH key required. Activity feeds keep you updated on progress.

More about collaboration

**Code review**

Comment on changes, track issues, compare branches, send pull requests and merge forks.

More about code review

**Reliable code hosting**

We spend all day and night making sure your repositories are **secure**, **backed up** and **always available**.

More about code hosting

**Open source collaboration**

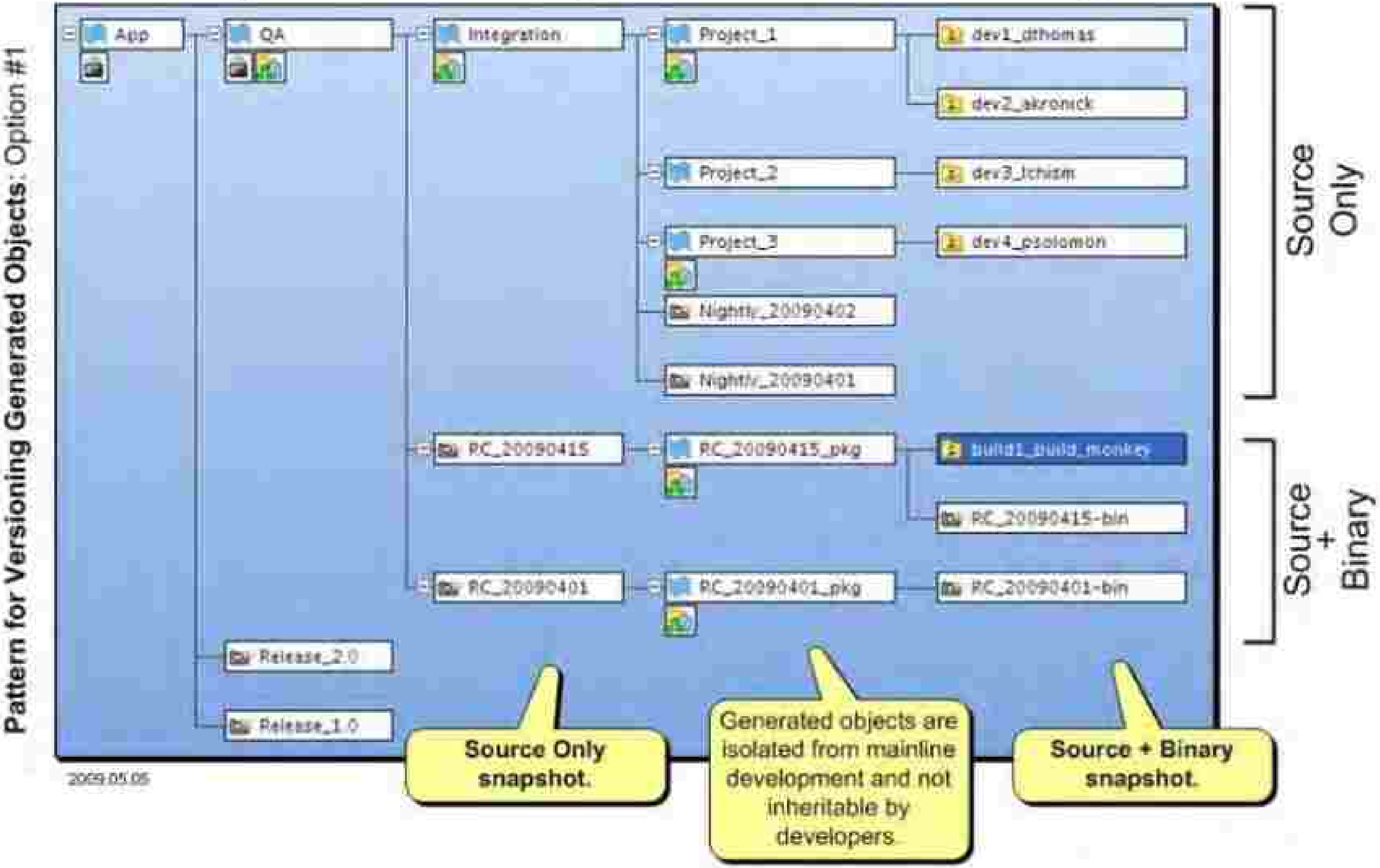
Participate in the most important open source community in the world today—online or at one of our meetups.

More about our community

© 2011 GitHub Inc. All rights reserved.

Powered by the Dedicated Servers and Cloud Computing of Rackspace Hosting®

# Why not share clinical /genomic data and model building in the ways currently used by the software industry (power of tracking workflows and versioning



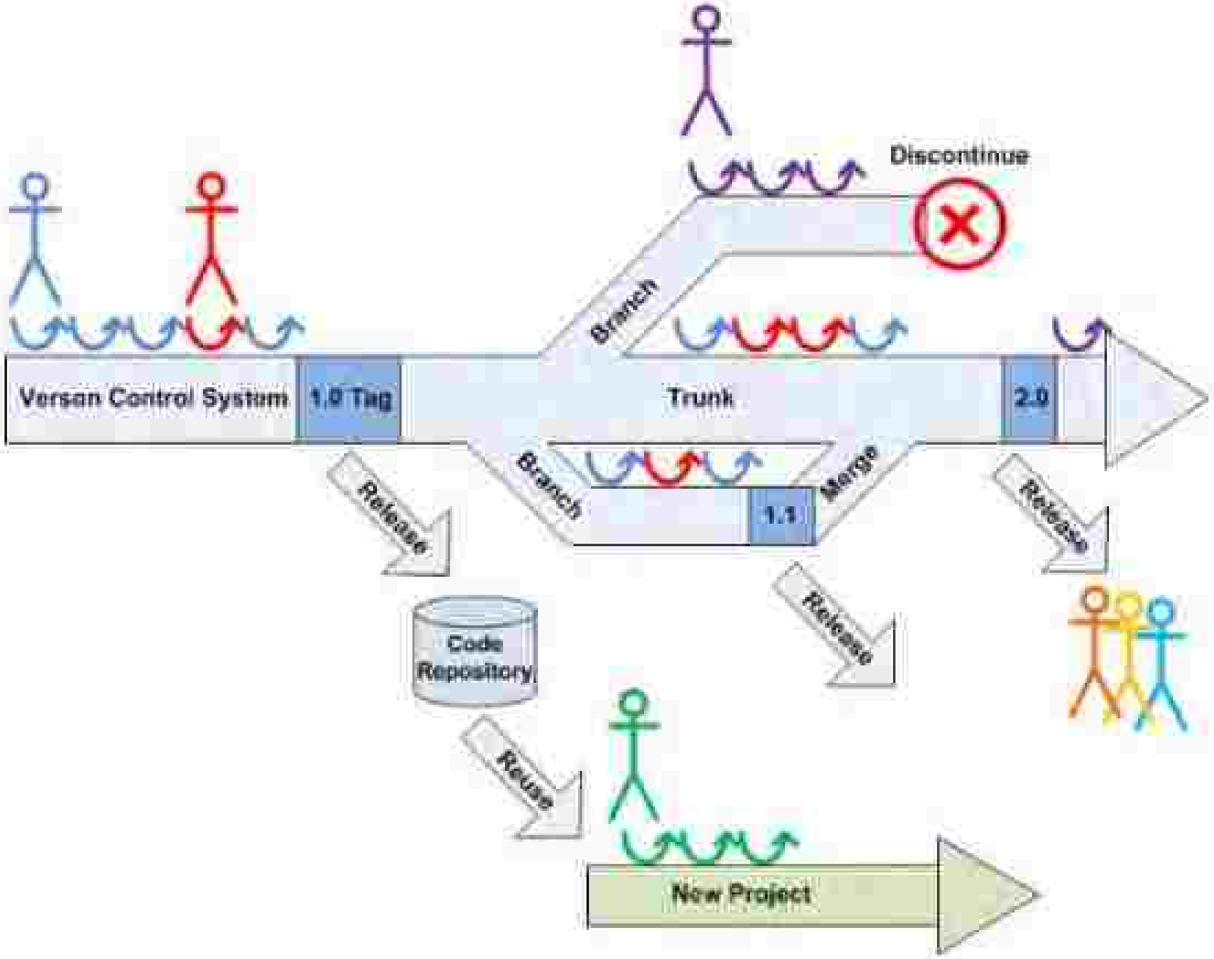
**Source Only snapshot.**

Generated objects are isolated from mainline development and not inheritable by developers.

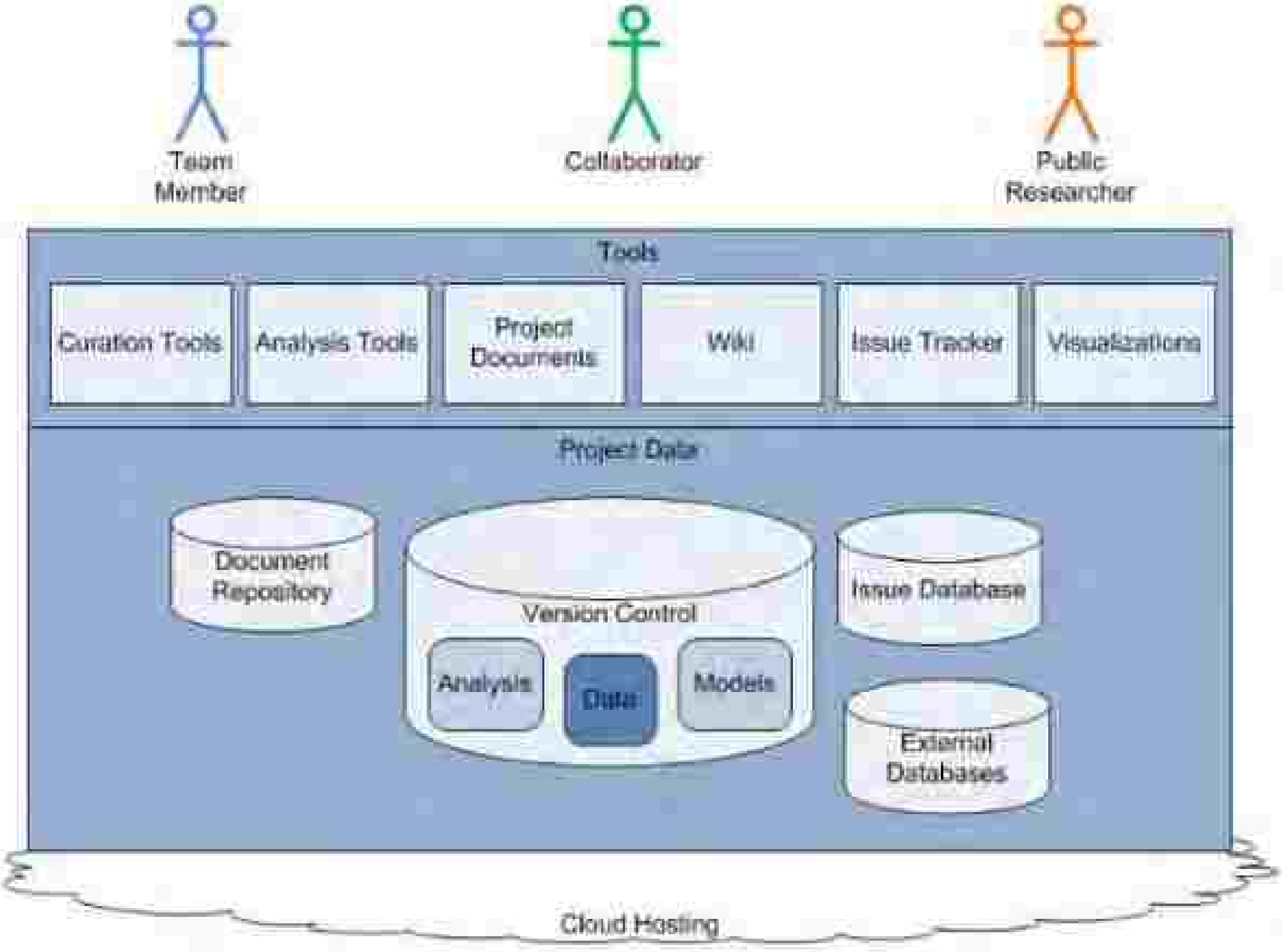
**Source + Binary snapshot.**

2009.05.05

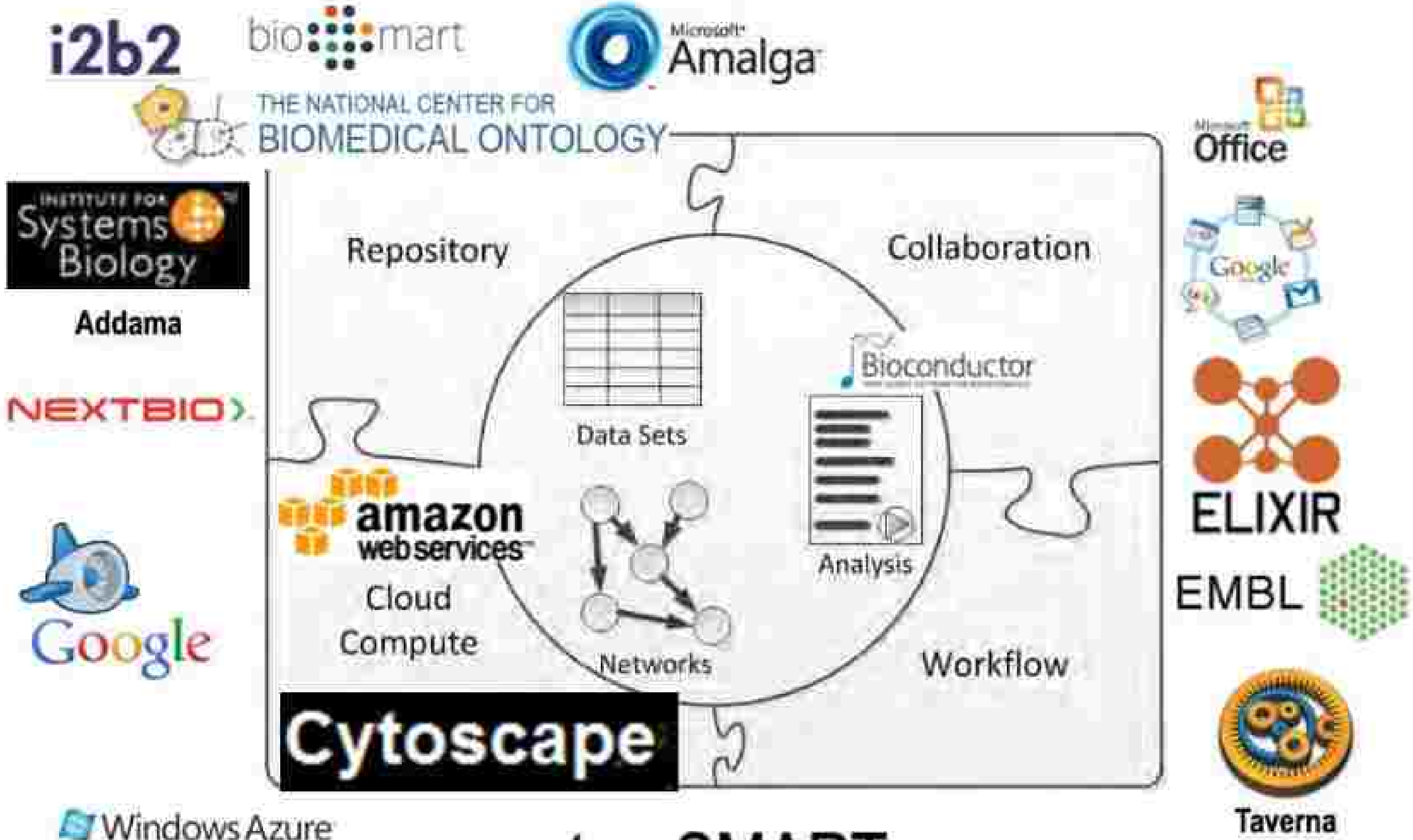
# Evolution of a Software Project



# Biology Tools Support Collaboration

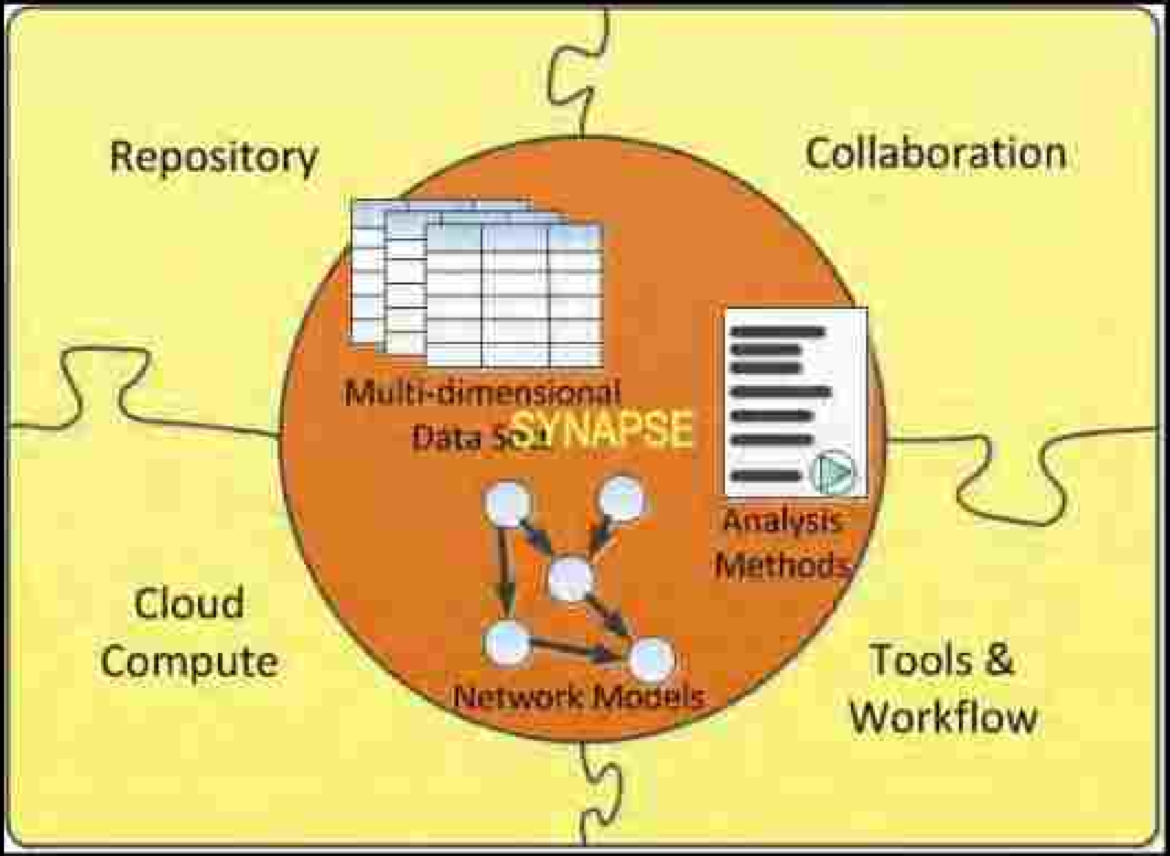


# Potential Supporting Technologies



tranSMART

# Platform for Modeling



# sage bionetworks synapse project

## Watch What I Do, Not What I Say

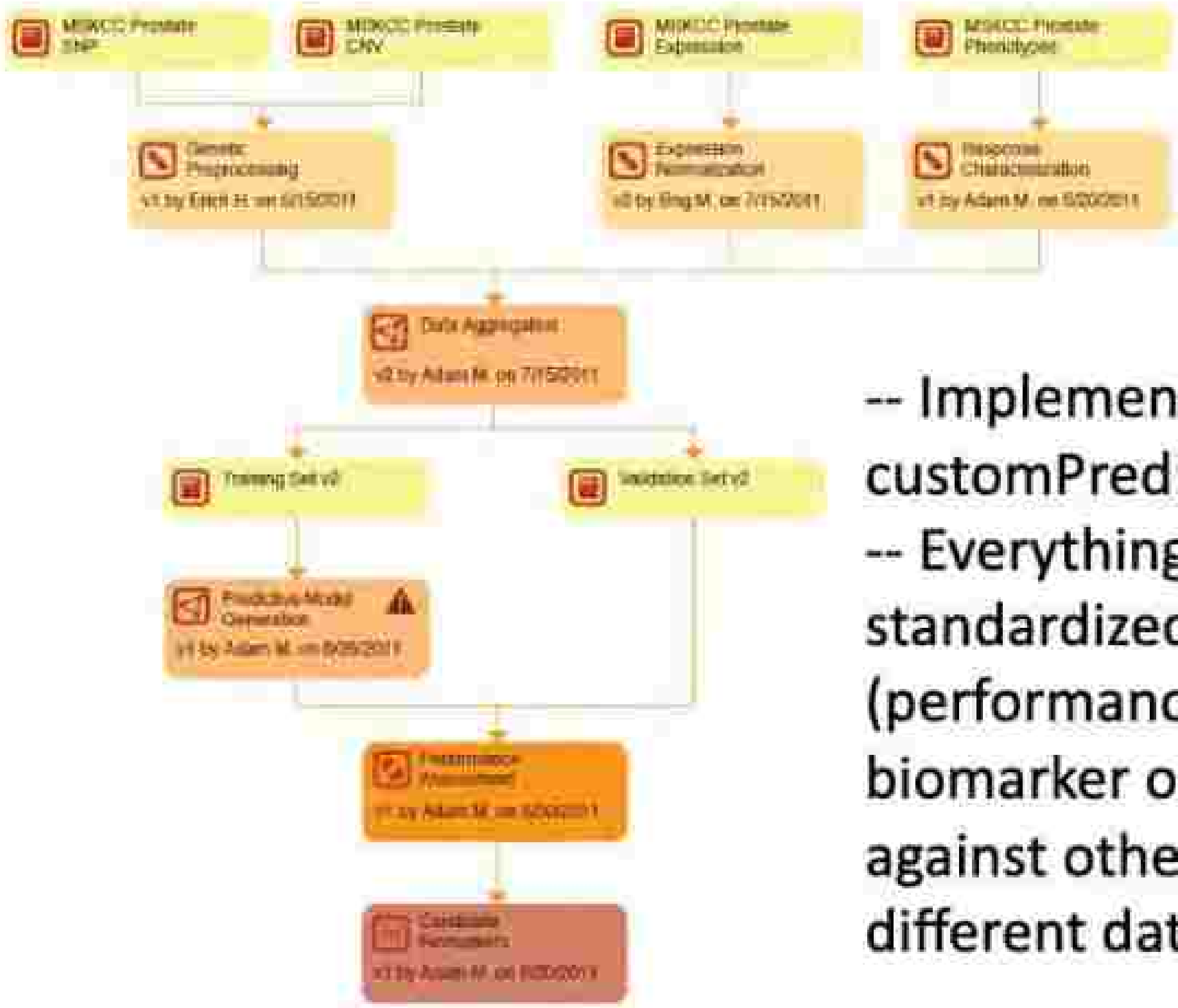
## Reduce, Reuse, Recycle

## Most of the People You Need to Work with Don't Work with You

## My Other Computer is Amazon

Sage BIONETWORKS

# Synapse machine learning infrastructure for method comparison

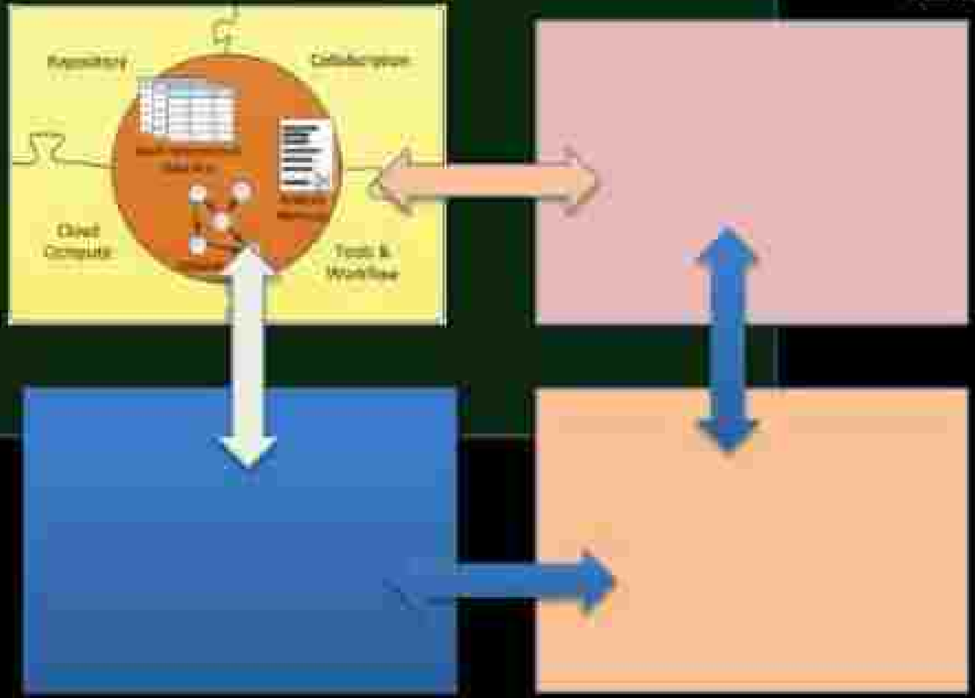

MSKCC Prostate SNP
MSKCC Prostate CNV
MSKCC Prostate Expression
MSKCC Prostate Phenotypes
Genetic Preprocessing
v1 by Erich H. on 6/15/2011
Expression Normalization
v2 by Brig M. on 7/15/2011
Response Characterization
v1 by Adam M. on 5/20/2011
Data Aggregation
v2 by Adam M. on 7/15/2011
Training Set v2
Validation Set v2
Predictive Model Generation
v1 by Adam M. on 8/26/2011
Performance Assessment
v1 by Adam M. on 8/20/2011
Candidate Biomarkers
v1 by Adam M. on 8/20/2011

-- Implement customTrain() and customPredict() functions

-- Everything else handled in standardized workflow (performance evaluation, biomarker outputs, evaluation against other methods, loading of different datasets, etc).

# INTEROPERABILITY

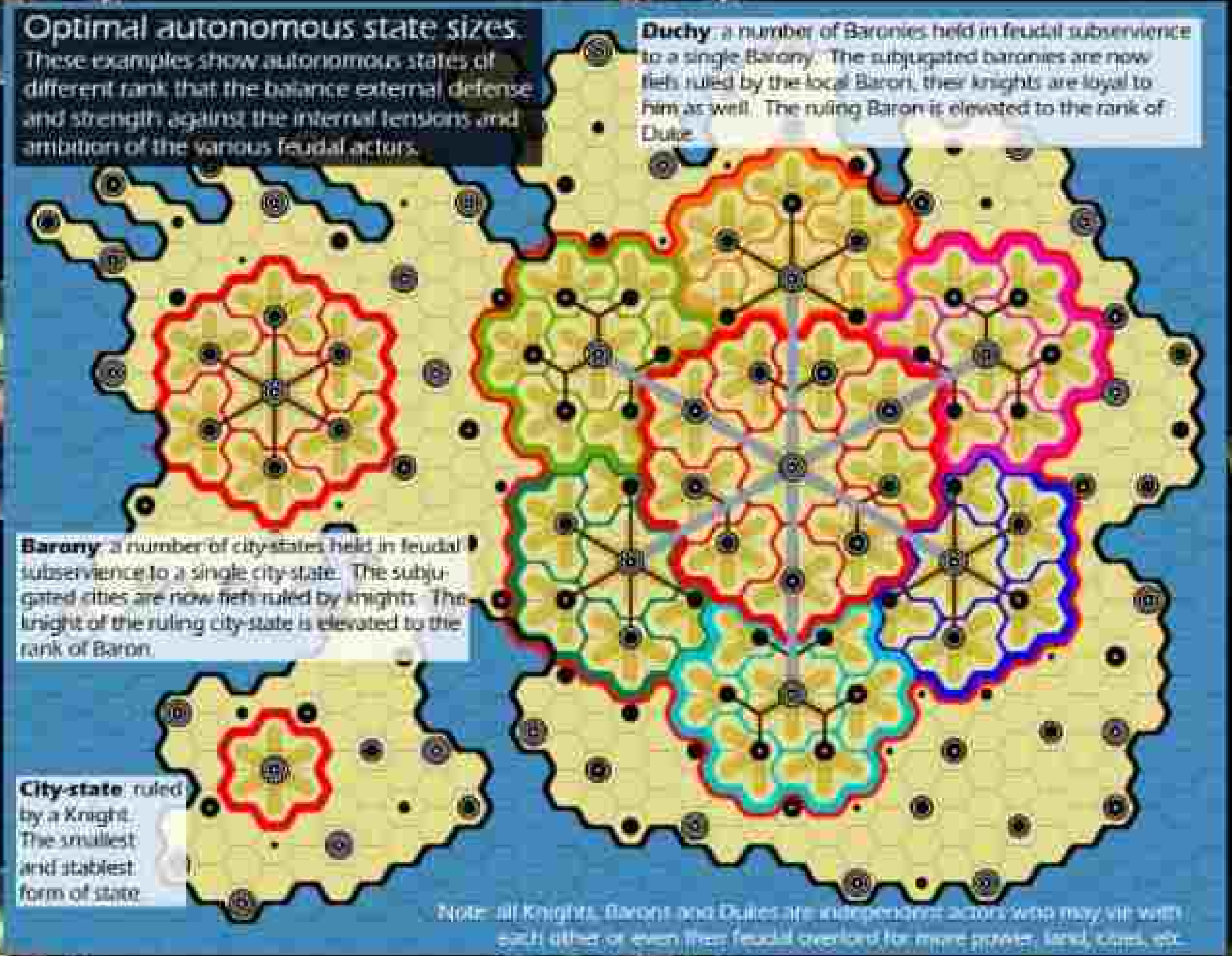
Genome Pattern
CYTOSCAPE
tranSMART
I2B2



# NOT JUST WHAT BUT HOW

hunter gathers- not sharing

# TENURE FEUDAL STATES



## Optimal autonomous state sizes.

These examples show autonomous states of different rank that the balance external defense and strength against the internal tensions and ambition of the various feudal actors.

**Duchy**: a number of Baronies held in feudal subservience to a single Barony. The subjugated baronies are now fiefs ruled by the local Baron, their knights are loyal to him as well. The ruling Baron is elevated to the rank of Duke.

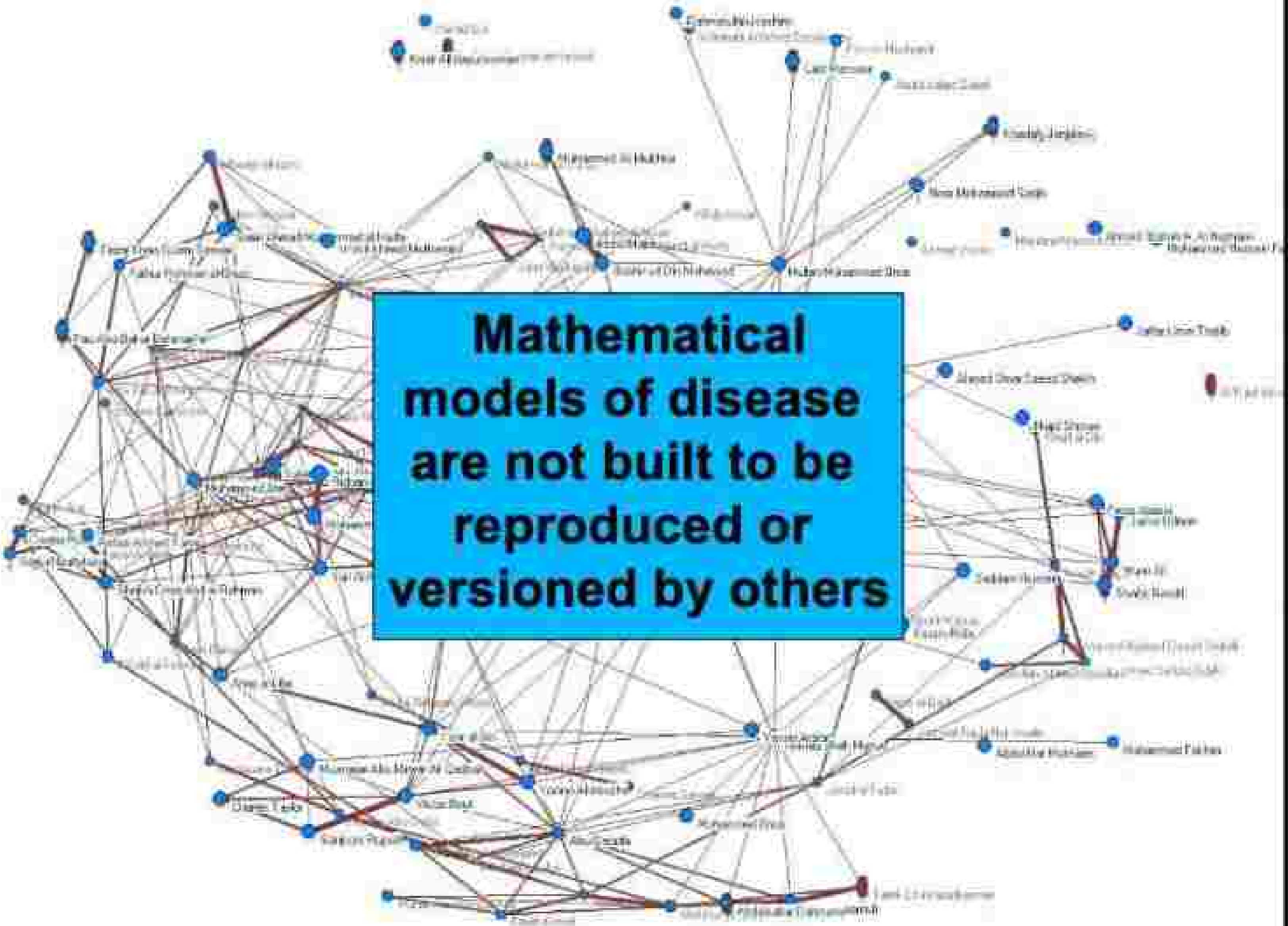
**Barony**: a number of city-states held in feudal subservience to a single city-state. The subjugated cities are now fiefs ruled by knights. The knight of the ruling city-state is elevated to the rank of Baron.

**City-state**: ruled by a Knight. The smallest and stablest form of state

Note: all Knights, Barons and Dukes are independent actors who may vie with each other or even their feudal overlord for more power, land, cities, etc.

Clinical/genomic data
are accessible but minimally usable

Little incentive to annotate and curate
data for other scientists to use



**Mathematical models of disease are not built to be reproduced or versioned by others**

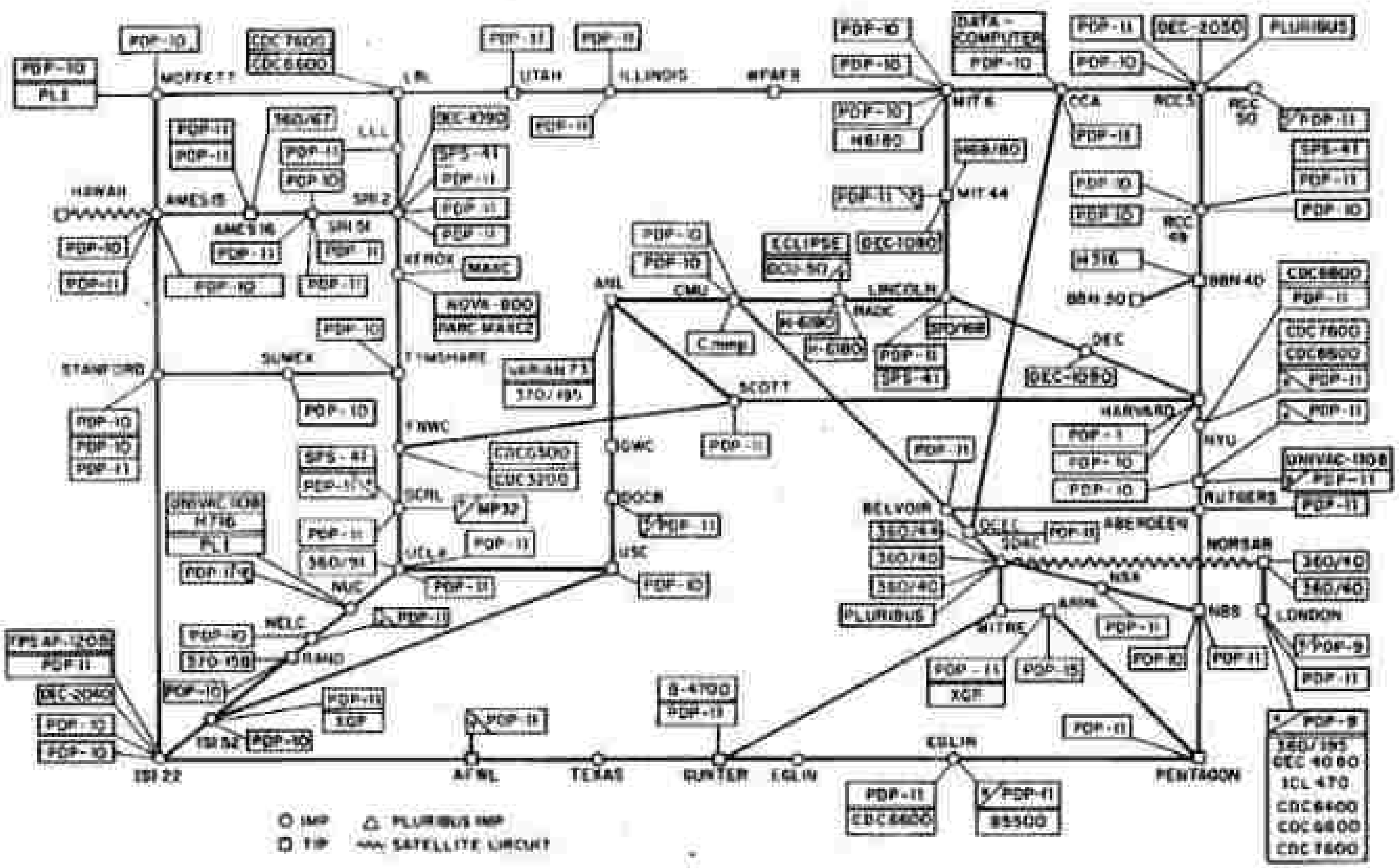
**Assumption that genetic alterations in human conditions should be owned**

ਪੋਥੀਆਂ ਗੁਰੂ ਹਰਿਸਹਾਇ ਵਾਲੀਆਂ ਵਿਚੋਂ ਇਕ ਭਾਗ ਦਾ ਆਰੰਭਕ ਸਫ਼ਾ ੴ ਸਤਿਨਾਮ—ਬਾਬਾ ਨਾਨਕ

## Lack of standard forms for sharing data and lack of forms for future rights and consentss

**Publication Bias- Where can we find the (negative) clinical data?**

ARPANET LOGICAL MAP, MARCH 1977



(PLEASE NOTE THAT WHILE THIS MAP SHOWS THE HOST POPULATION OF THE NETWORK ACCORDING TO THE BEST INFORMATION OBTAINABLE, NO CLAIM CAN BE MADE FOR ITS ACCURACY)

NAMES SHOWN ARE IMP NAMES, NOT (NECESSARILY) HOST NAMES

sharing as an adoption of common standards..

Clinical Genomics Privacy IP

# Six Pilots at Sage Bionetworks

CTCAP
Non-Responders
Arch2POCM
The Federation
Portable Legal Consent
Sage Congress Project

NEW TOOLS
PILOTS
PLATFORM
RULES GOVERN
NEW MAPS

# CTCAP

## Clinical Trial Comparator Arm Partnership "CTCAP"
## Strategic Opportunities For Regulatory Science Leadership and Action

FDA
September 27, 2011

Genetic Alliance

Sage Bionetworks

Johnson & Johnson

Pfizer

gsk GlaxoSmithKline

Roche

Lilly

MERCK

QUINTILES

medco

# Clinical Trial Comparator Arm Partnership (CTCAP)

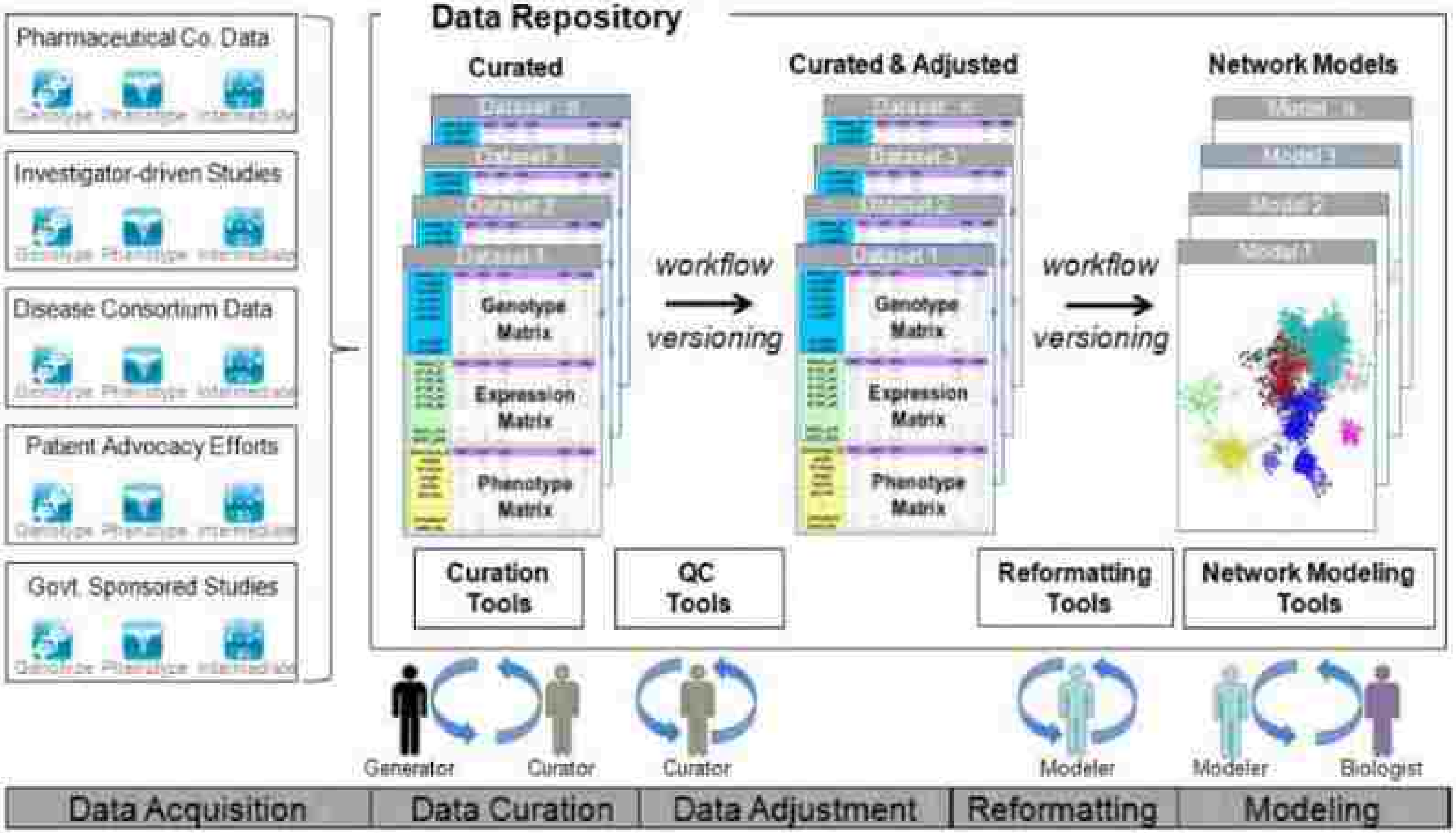
Partnering for Cures | A FasterCures Meeting

Bridging the Chasm Between Microscope and Marketplace

- **Description**: Collate, Annotate, Curate and Host Clinical Trial Data with Genomic Information from the Comparator Arms of Industry and Foundation Sponsored Clinical Trials: Building a Site for Sharing Data and Models to evolve better Disease Maps.
- **Public-Private Partnership** of leading pharmaceutical companies, clinical trial groups and researchers.
- **Neutral Conveners**: Sage Bionetworks and Genetic Alliance [nonprofits].
- **Initiative to share existing trial data** (molecular and clinical) from non-proprietary comparator and placebo arms to create powerful new tool for drug development.

Started Sept 2010

# Shared clinical/genomic data sharing and analysis will maximize clinical impact and enable discovery



Pharmaceutical Co. Data

Investigator-driven Studies

Disease Consortium Data

Patient Advocacy Efforts

Govt. Sponsored Studies

Data Repository

Curated

Curated & Adjusted

Network Models

Genotype Matrix

Expression Matrix

Phenotype Matrix

*workflow* → *versioning*

Curation Tools

QC Tools

Reformatting Tools

Network Modeling Tools

Generator

Curator

Curator

Modeler

Modeler

Biologist

Data Acquisition | Data Curation | Data Adjustment | Reformatting | Modeling

# Non-Responders Project

To identify Non-Responders to approved Oncology drug regimens in order to improve outcomes, spare patients unnecessary toxicities from treatments that have no benefit to them, and reduce healthcare costs

# The Non-Responder Cancer Project Leadership Team

**Stephen Friend, MD, PhD**
President and Co-Founder of Sage Bionetworks, Head of Merck Oncology 01-08, Founder of Rosetta Inpharmatics 97-01, co-Founder of the Seattle Project

**Todd Golub, MD**
Founding Director Cancer Biology Program Broad Institute, Charles Dana Investigator Dana-Farber Cancer Institute, Professor of Pediatrics Harvard Medical School, Investigator, Howard Hughes Medical Institute

**Garry Nolan, PhD**
Professor, Baxter Laboratory of Stem Cell Biology, Department of Microbiology and Immunology, Stanford University Director, Proteomics Center at Stanford University

**Richard Schilsky, MD**
Chief, Hematology- Oncology, Deputy Director, Comprehensive Cancer Center, University of Chicago; Chair, National Cancer Institute Board of Scientific Advisors; past-President ASCO, past Chairman CALGB clinical trials group

# The Non-Responder Project is an international initiative with funding for 6 initial cancers anticipated from both the public and private sectors

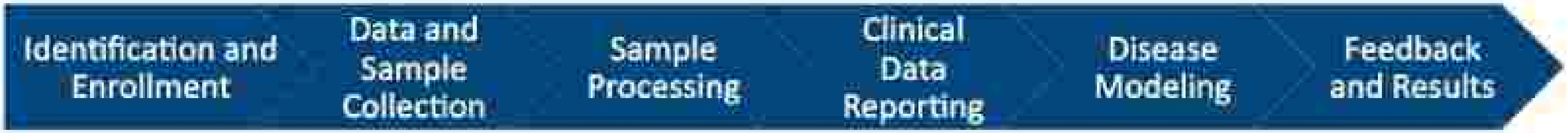
| | | | | | | |
|---|---|---|---|---|---|---|
| GEOGRAPHY | United States | | | | China | |
| TARGET CANCER | Ovarian | Renal | Breast | AML | Colon | Lung |
| FUNDING SOURCE | Seeking private sector and philanthropic funding for prospective studies | | | Retrospective study; likely to be funded by the Federal Government | Funded by the Chinese government and private sector partners | |

# For each tumor-type, the non-responder project will follow a common workflow, with patient identification and sample collection the most variable across studies

**Non-Responder Project Workflow**

Identification and enrollment, and data and sample collection may differ by tumor-type

The remaining parts of the study will be largely similar, and potentially shared, across all projects



Identification and Enrollment

Data and Sample Collection

Sample Processing

Clinical Data Reporting

Disease Modeling

Feedback and Results

Payment and Reimbursement

Project Management

# A consortium of collaborators has been constructed to execute the non-responder project

**Physicians & AMCs**

CEDARS-SINAI

Mount Sinai The Mount Sinai Medical Center

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center

DANA-FARBER
CANCER INSTITUTE

MAYO CLINIC

Memorial Sloan-Kettering
Cancer Center

STANFORD
SCHOOL OF MEDICINE

MASSACHUSETTS
GENERAL HOSPITAL
1811 - 2011

**Patient Advocacy Groups**

THE CLEARITY FOUNDATION

army of women

# A consortium of collaborators has been constructed to execute the non-responder project (continued)

| | | |
|---|---|---|
| Patient Consent | Genetic Alliance | |
| Pathology | THE UNIVERSITY OF TEXAS<br>MD Anderson<br>~~Cancer~~ Center | Memorial Sloan-Kettering<br>Cancer Center |
| Genome Sequencing | TBD | |
| Core Bioinformatics | TGen | UNIVERSITY OF CALIFORNIA<br>SANTA CRUZ |
| Analysis and Disease Modeling | Sage<br>BIONETWORKS | BROAD<br>INSTITUTE |

# Arch2POCM

## Restructuring Drug Discovery

## How to potentially De-Risk High-Risk Therapeutic Areas

# What is the problem?

- Regulatory hurdles too high?
- Low hanging fruit picked?
- Payers unwilling to pay?
- Genome has not delivered?
- Valley of death?
- Companies not large enough to execute on strategy?
- Internal research costs too high?
- Clinical trials in developed countries too expensive?

In fact, all are true but none is the real problem

# What is the problem?

We need to rebuild the drug discovery process so that we better understand disease biology before testing proprietary compounds on sick patients

THERAPEUTICS DISCOVERY

# The Precompetitive Space: Time to Move the Yardsticks

Thea Norman,[1] Aled Edwards,[2] Chas Bountra,[3] Stephen Friend[4]*

Industry, government, patient advocacy groups, public funders, and academic thought leaders met in Toronto, Canada, to set into motion an initiative that addresses some of the scientific and organizational challenges of modern therapeutics discovery. What emerged from the meeting was a public-private partnership that seeks to establish proof of clinical mechanism (POCM) for selected "pioneer" disease targets using lead compounds—all accomplished in the precompetitive space. The group will reconvene in April 2011 to create a business plan that specifies the generation of two positive POCM results per year.

2011 m
which
formed
realm o
the bold
Februar
are enac
key tho
ment, p
ing age
followir
morrow
How m
and how
With th
try pro
more t
this ye
sales of
tives to
industry
Atte
Summit
[illegible]

MEETING REPORT

CROWDSOURCING

# Leveraging Crowdsourcing to Facilitate the Discovery of New Medicines

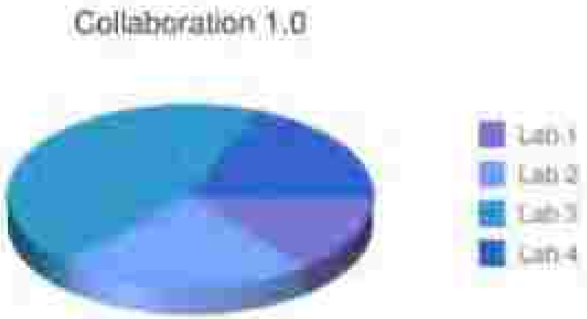
Thea C. Norman,[1] Chas Bountra,[2] Aled M. Edwards,[3] Keith R. Yamamoto,[4] Stephen H. Friend[5]*

Gloomy predictions about the future of pharma have forced the industry to investigate alternative models of drug discovery. Public-private partnerships (PPPs) have the potential to revitalize the discovery and development of first-in-class therapeutics. The new PPP Arch2POCM hopes to foster biomedical innovation through precompetitive validation of pioneer therapeutic targets for human diseases. In this meeting report, we capture the most exciting insights garnered from the April 2011 Arch2POCM conference.

When useful knowledge exists in companies of all sizes and also in universities, non-profits and individual minds, it makes sense to orient your innovation efforts to accessing, building upon and integrating that external knowledge into useful products and services.

# The Federation

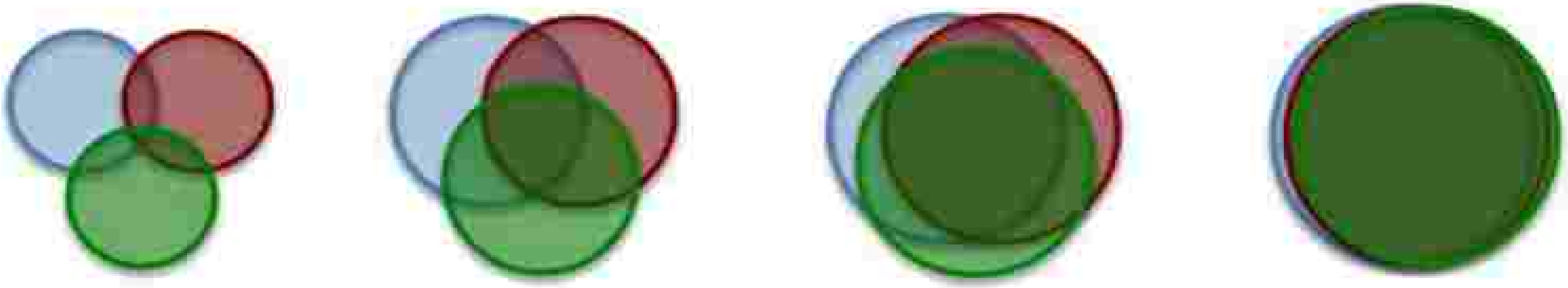
# How can we accelerate the pace of scientific discovery?



Ways to move beyond "traditional" collaborations?

Intra-lab vs Inter-lab Communication

Colrain/ Industrial PPPs Academic Unions



## THE FEDERATION

Butte Califano Friend Ideker Schadt

vs

# Rules of the game:
## transparency & trust

- Shared data tools models and prepublications
- Conflict of interests
- Intellectual property
- Authorship

# sage federation:
## human aging project

Sage

Justin Guinney
Stephen Friend*

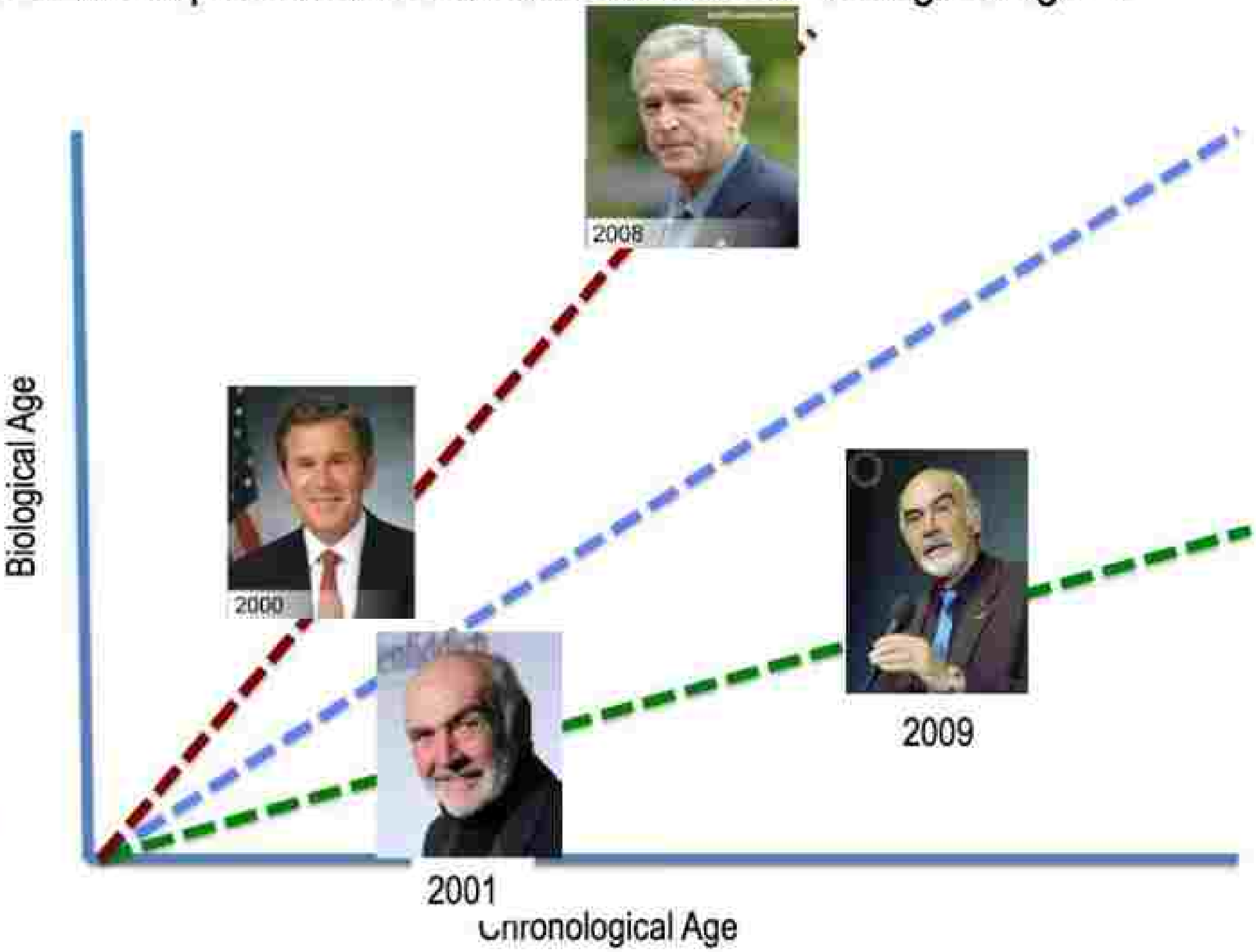
UC San Diego
School of Medicine

Greg Hannum
Januz Dutkowski
Trey Ideker*
Kang Zhang*

Columbia University

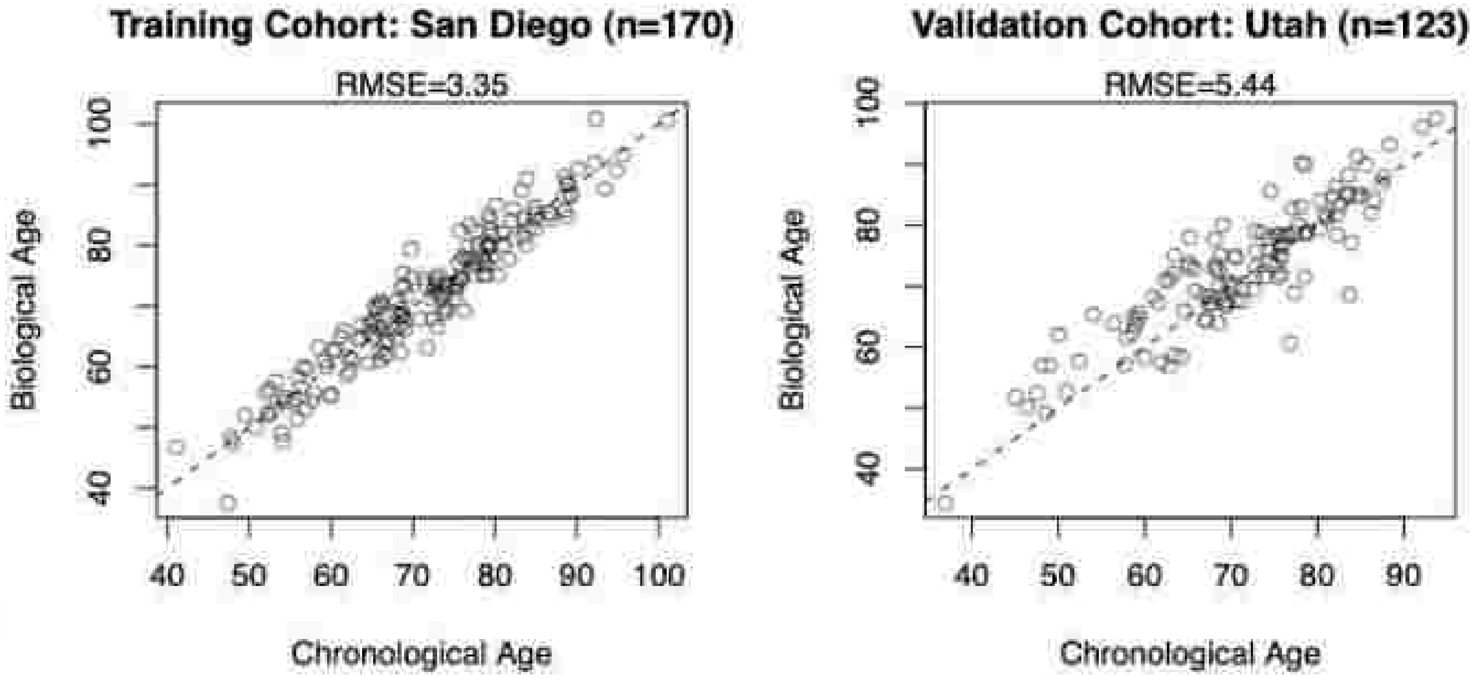
Mariano Alvarez
Celine Lefebrev
Andrea Califano*

## sage federation:
## what is the impact of disease/environment on “biological age” ?



# human aging:
## predicting bioage using whole blood methylation

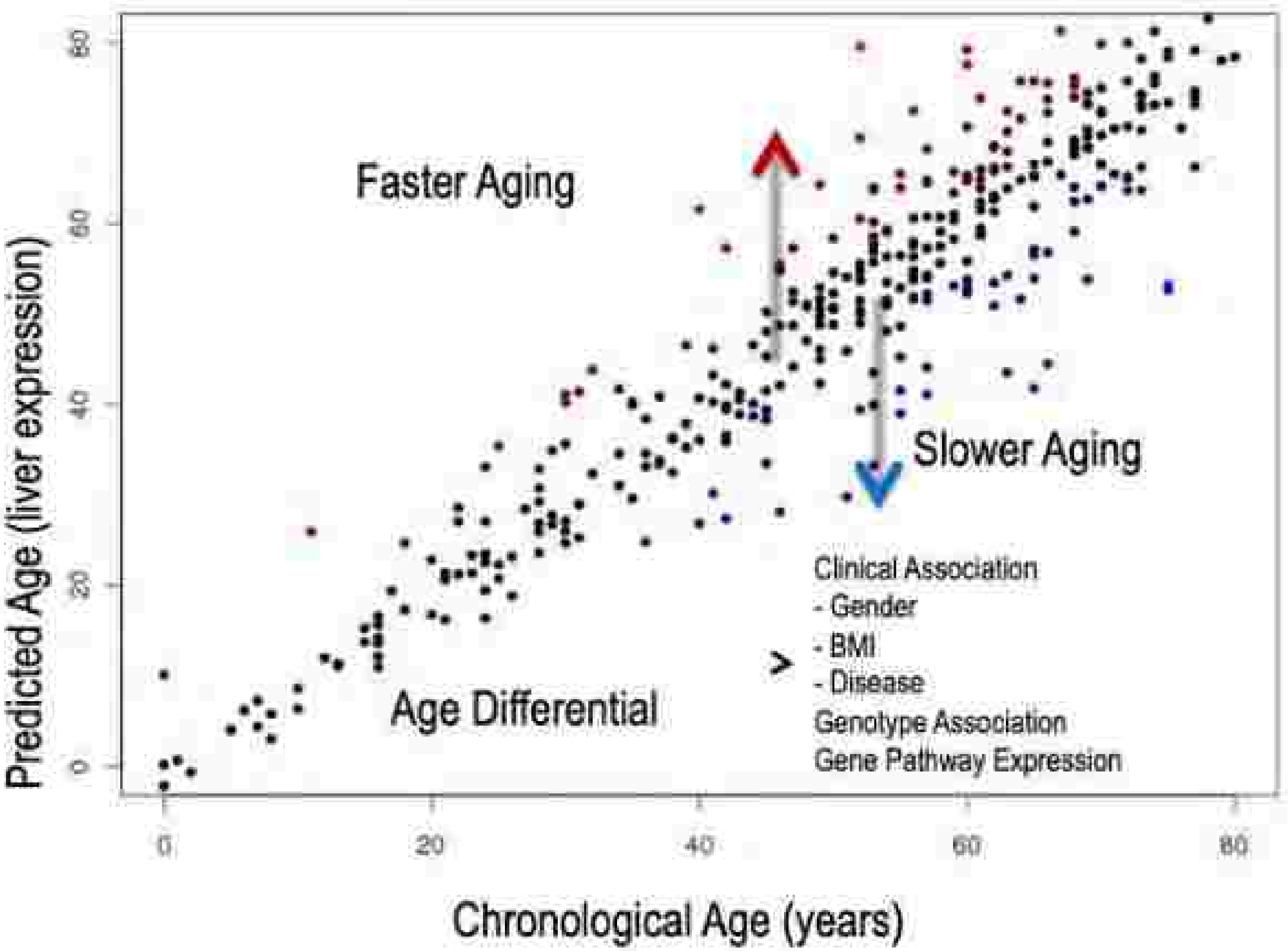
- Independent training (n=170) and validation (n=123) Caucasian cohorts
- 450k Illumina methylation array
- Exom sequencing
- Clinical phenotypes: Type II diabetes, BMI, gender...



## sage federation:
## model of biological age

$$\text{Bioage} = f(M) = Age + \sum_J \alpha_j C_j + \epsilon$$

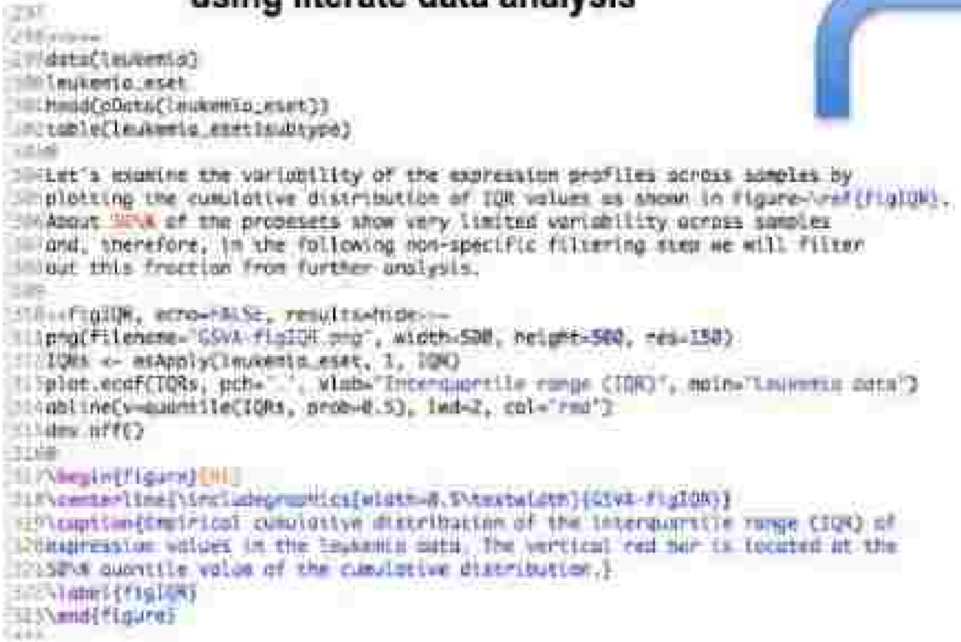
$$\text{Differential Bioage} = f(M) - Age = \sum_J \alpha_j C_j + \epsilon$$



# Reproducible science==shareable science

## Sweave: combines programmatic analysis with narrative

### Dynamic generation of statistical reports using literate data analysis



```
<<>>=
data(leukemia)
leukemia_eset
head(pData(leukemia_eset))
table(leukemia_eset$subtype)
@
Let's examine the variability of the expression profiles across samples by
plotting the cumulative distribution of IQR values as shown in Figure~\ref{figIQR}.
About 50\% of the probesets show very limited variability across samples
and, therefore, in the following non-specific filtering step we will filter
out this fraction from further analysis.
@

<<figIQR, echo=FALSE, results=hide>>=
png(filename="GSVA-figIQR.png", width=500, height=500, res=150)
IQRs <- esApply(leukemia_eset, 1, IQR)
plot.ecdf(IQRs, pch=".", xlab="Interquartile range (IQR)", main="Leukemia data")
abline(v=quantile(IQRs, prob=0.5), lwd=2, col="red")
dev.off()
@
\begin{figure}[h]
\centerline{\includegraphics[width=0.5\textwidth]{GSVA-figIQR}}
\caption{Empirical cumulative distribution of the interquartile range (IQR) of
expression values in the leukemia data. The vertical red bar is located at the
50\% quantile value of the cumulative distribution.}
\label{figIQR}
\end{figure}
<<>>=
```

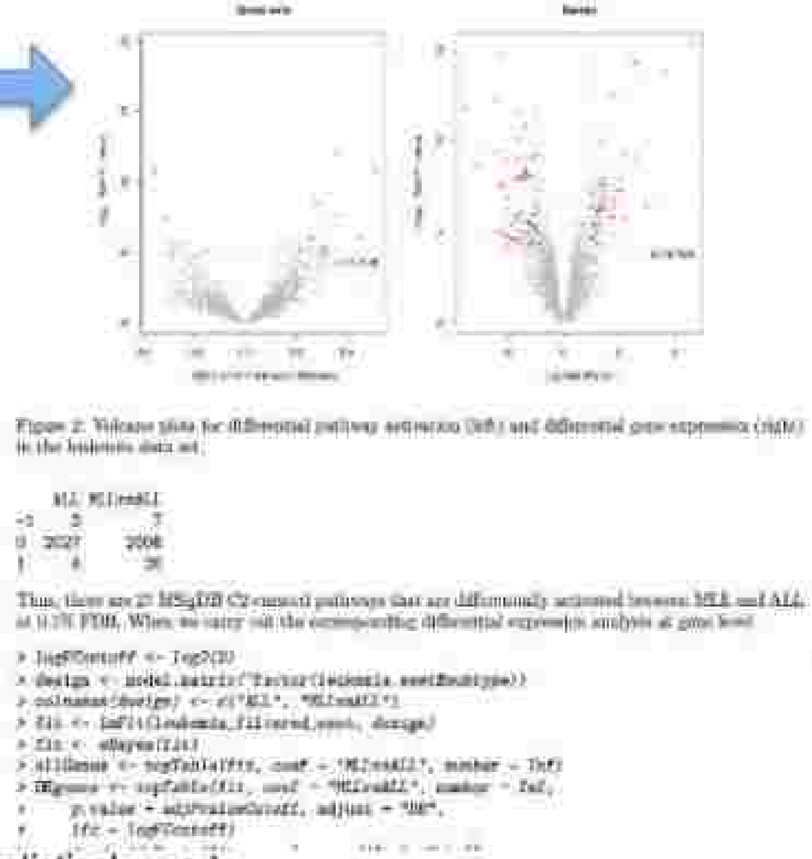


Figure 2: Volcano plots for differential pathway activation (left) and differential gene expression (right) in the leukemia data set.

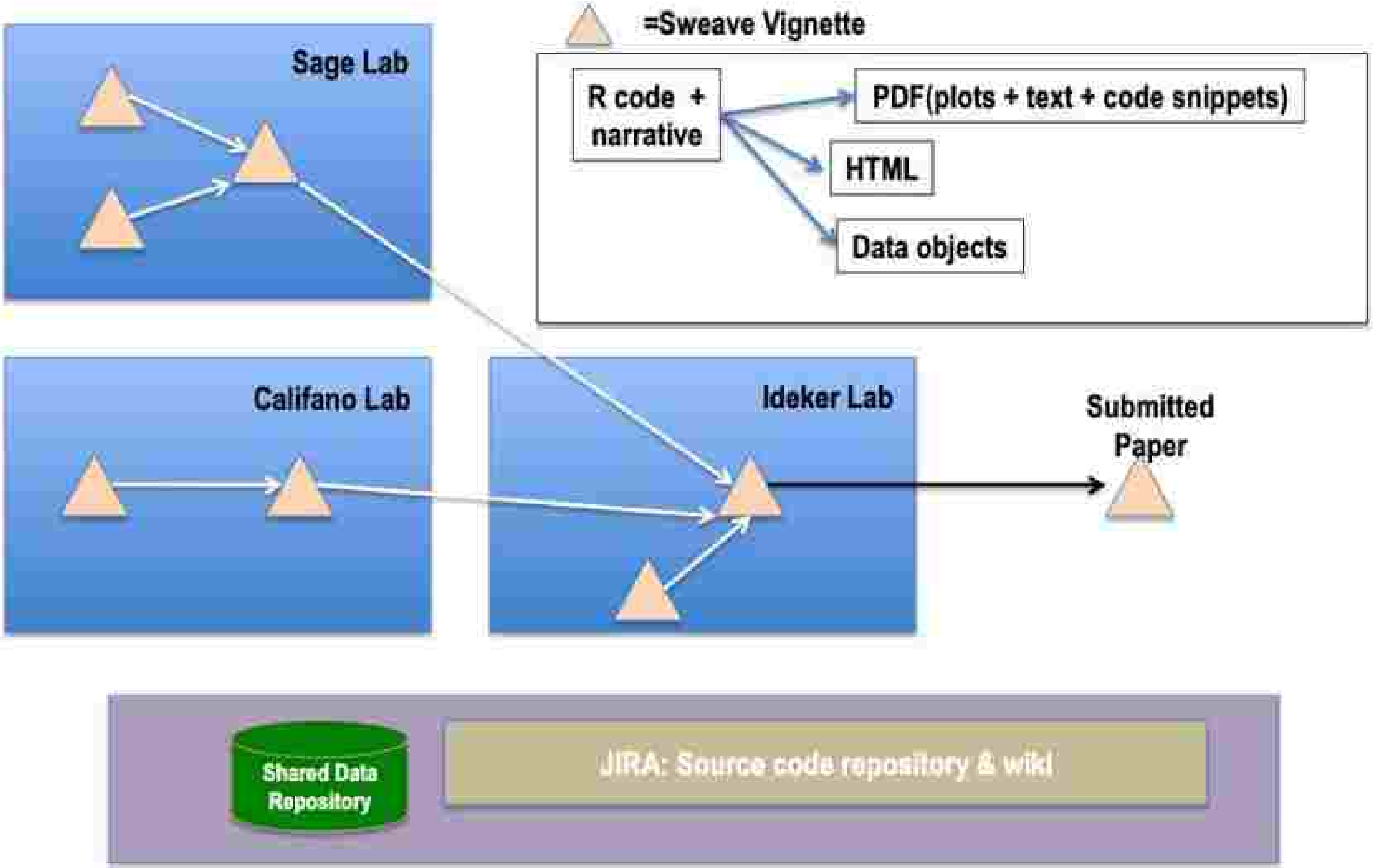
```
   ALL MLLrearALL
-1    5          7
0  2027       2008
1     6         26
```

Thus, there are 27 MSigDB C2 curated pathways that are differentially activated between MLL and ALL at 0.1% FDR. When we carry out the corresponding differential expression analysis at gene level

```
> logFCcutoff <- log2(3)
> design <- model.matrix(~factor(leukemia_eset$subtype))
> colnames(design) <- c("ALL", "MLLrearALL")
> fit <- lmFit(leukemia_filtered_eset, design)
> fit <- eBayes(fit)
> allGenes <- topTable(fit, coef = "MLLrearALL", number = Inf)
> DEgenes <- topTable(fit, coef = "MLLrearALL", number = Inf,
+     p.value = adjPvalueCutoff, adjust = "BH",
+     lfc = logFCcutoff)
```

Sweave.Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz,editors, Compstat 2002 – Proceedings in Computational Statistics,pages 575-580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9

# Federated Aging Project :
# Combining analysis + narrative



=Sweave Vignette

R code + narrative

PDF(plots + text + code snippets)

HTML

Data objects

Sage Lab

Califano Lab

Ideker Lab

Submitted Paper

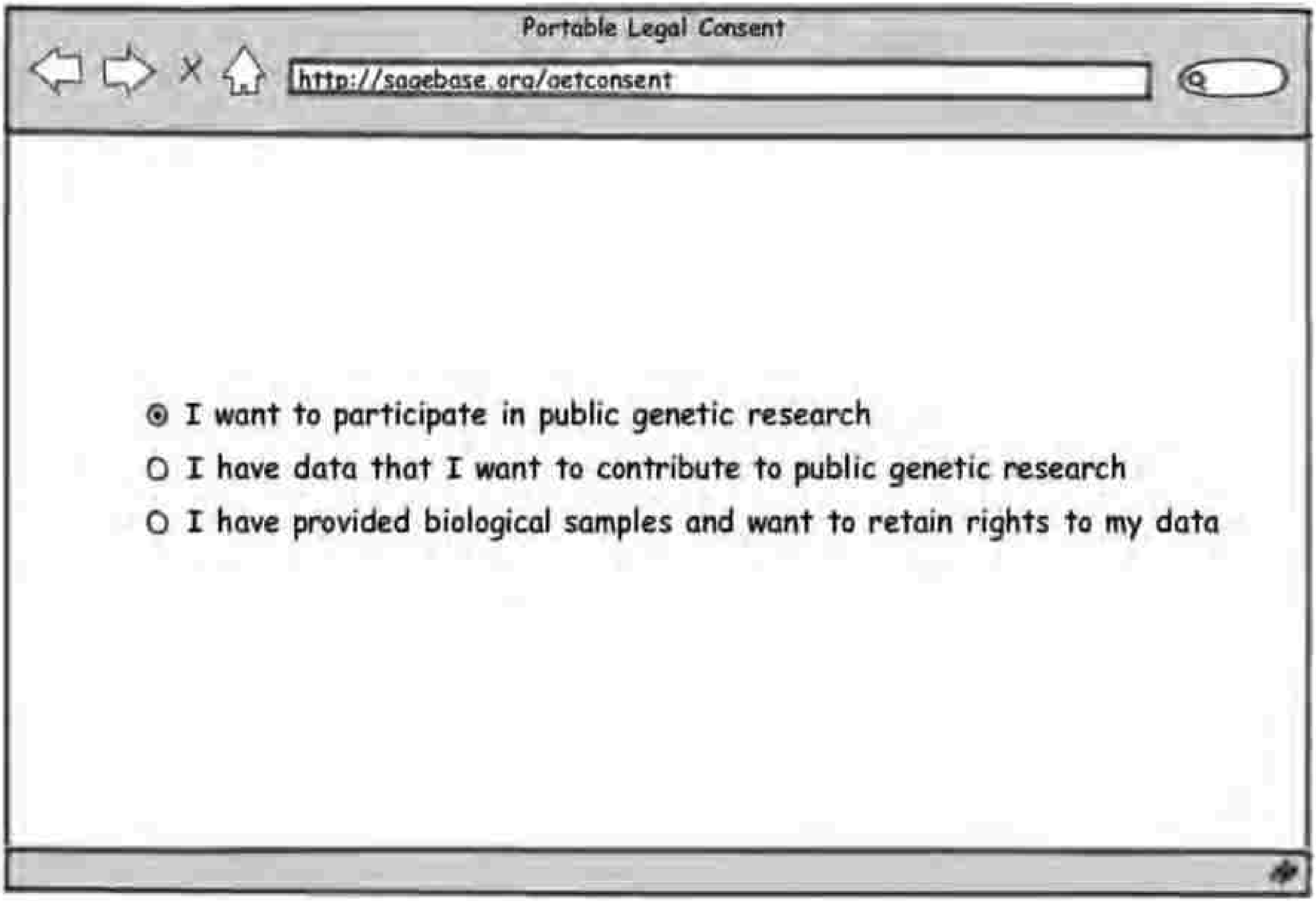
Shared Data Repository

JIRA: Source code repository & wiki

# Portable Legal Consent

(Activating Patients)

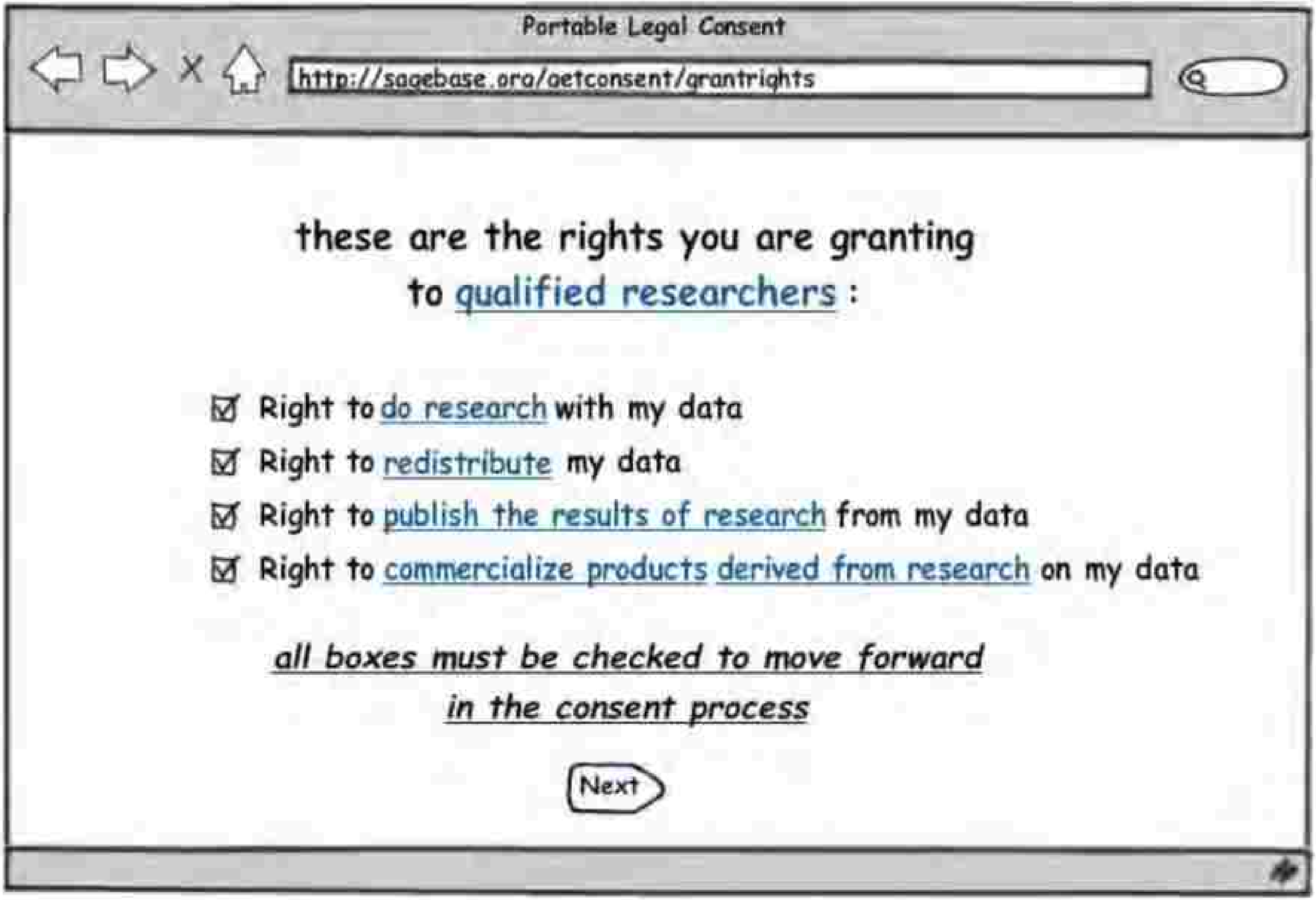
John Wilbanks



◉ I want to participate in public genetic research

○ I have data that I want to contribute to public genetic research

○ I have provided biological samples and want to retain rights to my data



http://sagebase.org/getconsent/grantrights

these are the rights you are granting to qualified researchers :

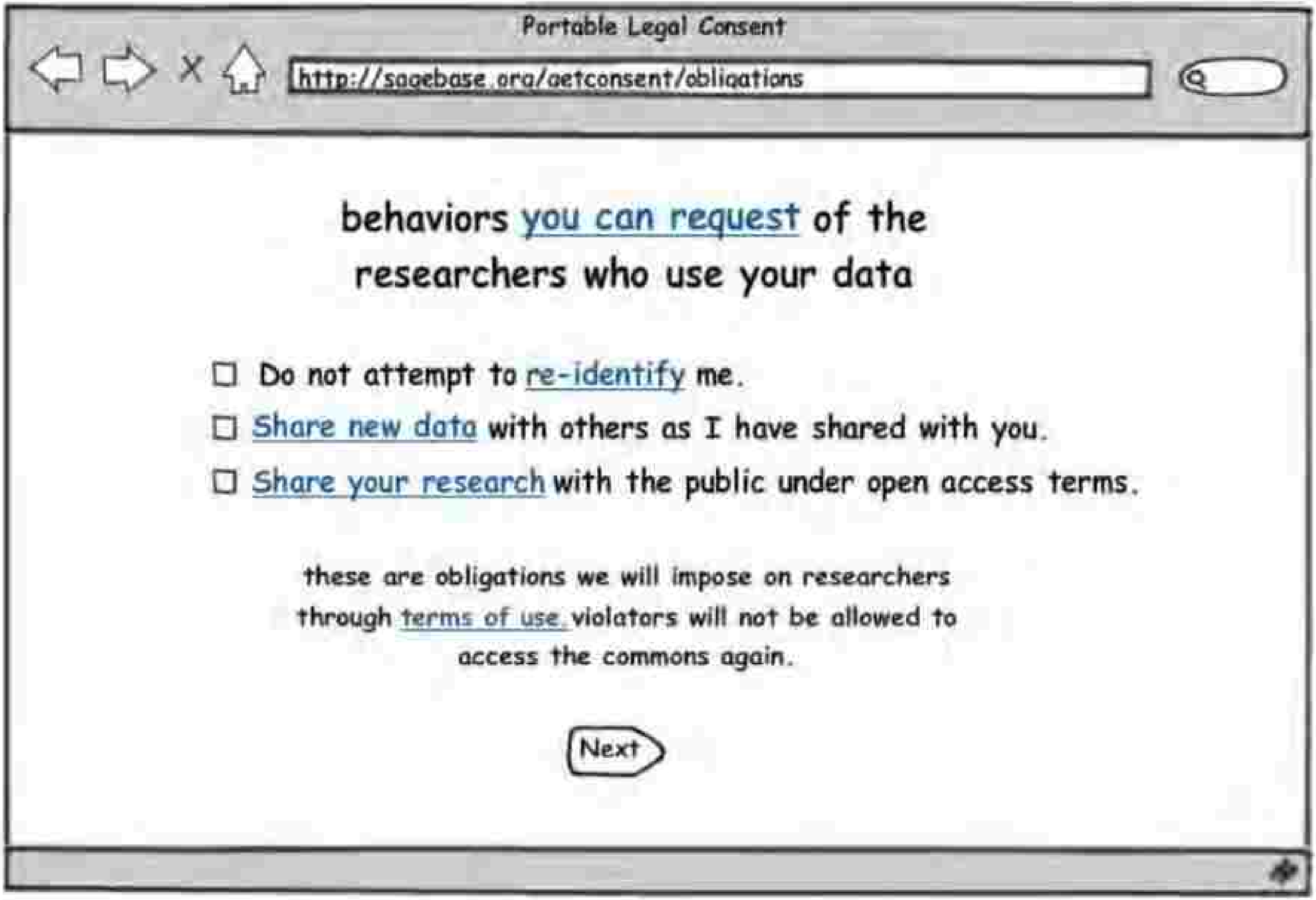
☑ Right to do research with my data

☑ Right to redistribute my data

☑ Right to publish the results of research from my data

☑ Right to commercialize products derived from research on my data

*all boxes must be checked to move forward in the consent process*



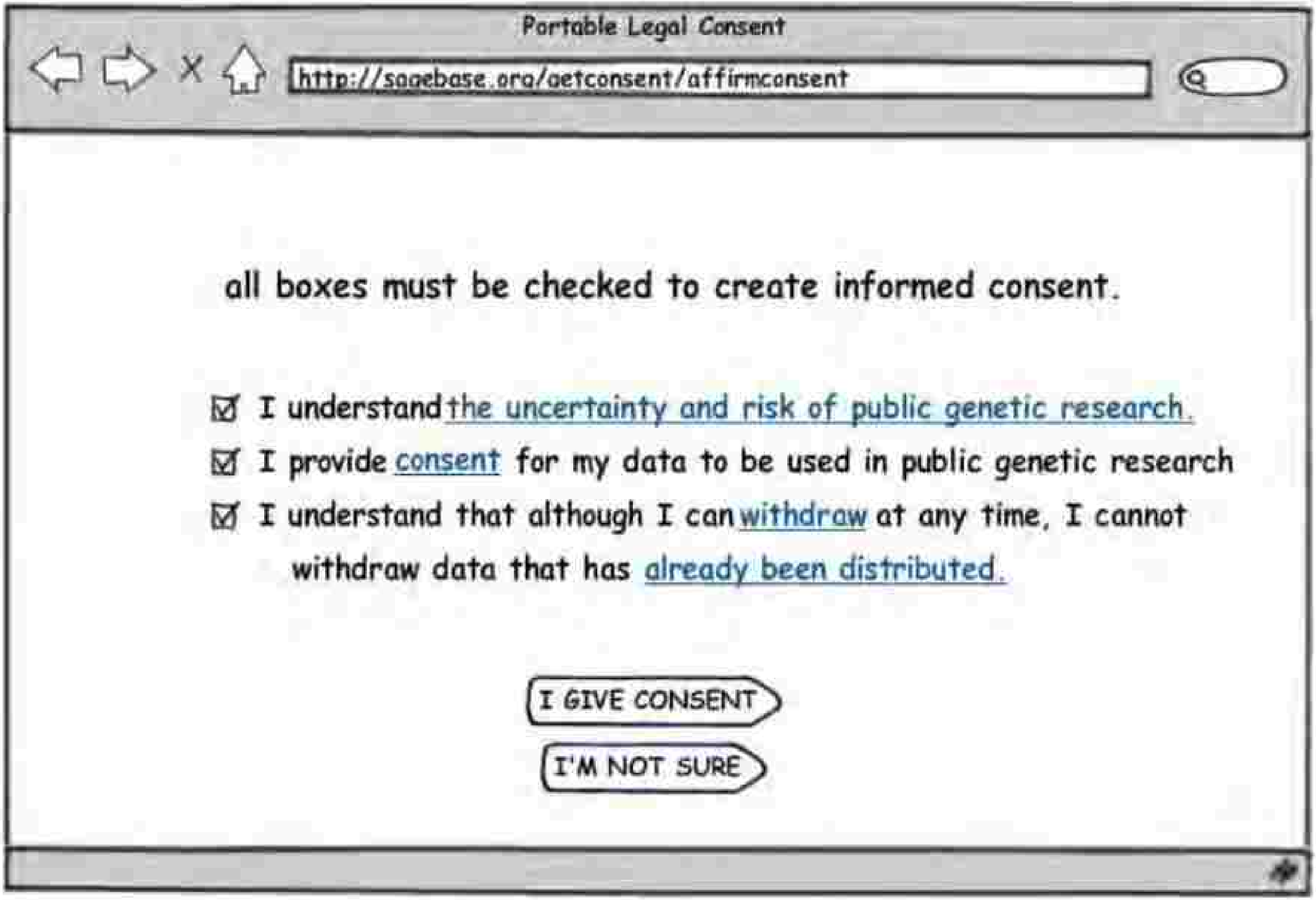
behaviors you can request of the researchers who use your data

☐ Do not attempt to re-identify me.

☐ Share new data with others as I have shared with you.

☐ Share your research with the public under open access terms.

these are obligations we will impose on researchers through terms of use. violators will not be allowed to access the commons again.



all boxes must be checked to create informed consent.

☑ I understand the uncertainty and risk of public genetic research.

☑ I provide consent for my data to be used in public genetic research

☑ I understand that although I can withdraw at any time, I cannot withdraw data that has already been distributed.

# Sage Congress Project
April 20 2012

RA

Parkinson's

Asthma

(Responders Competitions)

Why not use data intensive science
to build models of disease

Organizational Structures and Tools

How not What

Six Pilots

Opportunities