(Access to large scale data networks)

BETTER MAPS OF DISEASE

NOT JUST WHAT BUT HOW

BUILDING A COMMONS FOR EVOLVING GENERATIVE MODELS OF DISEASE

Academia

Biotech

Industry

Sage Bionetworks

Non-Profit

Sharing data and Building integrative disease models

Existing approaches and issues

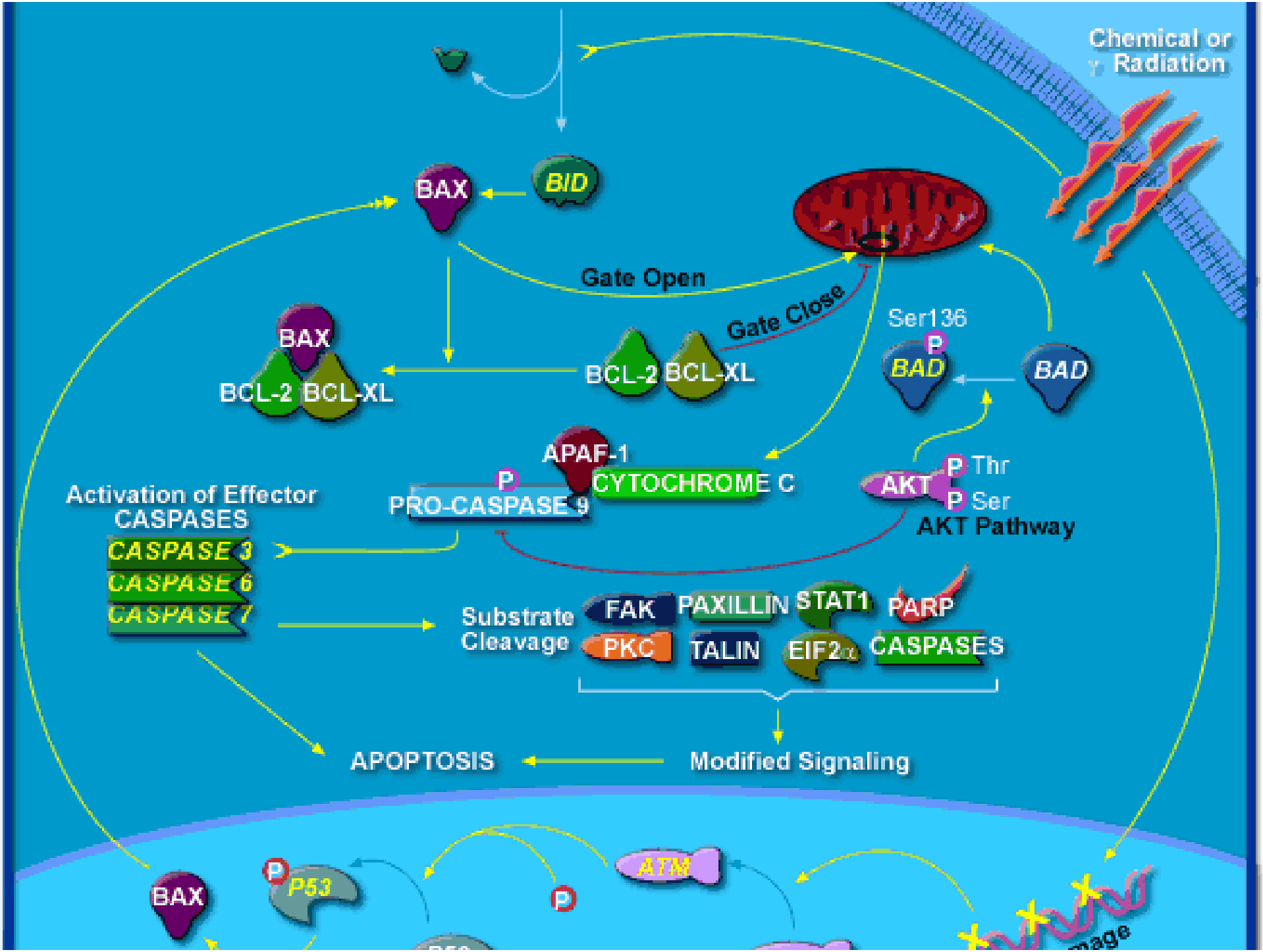
Cancer- 75% of drugs approved- ”standards of care” lack significant impact

25,000 components with 3,269 associated with disease-yet only hundreds targeted for therapies

Current costs for drug approval- ~$1Billion – 5 -10 years

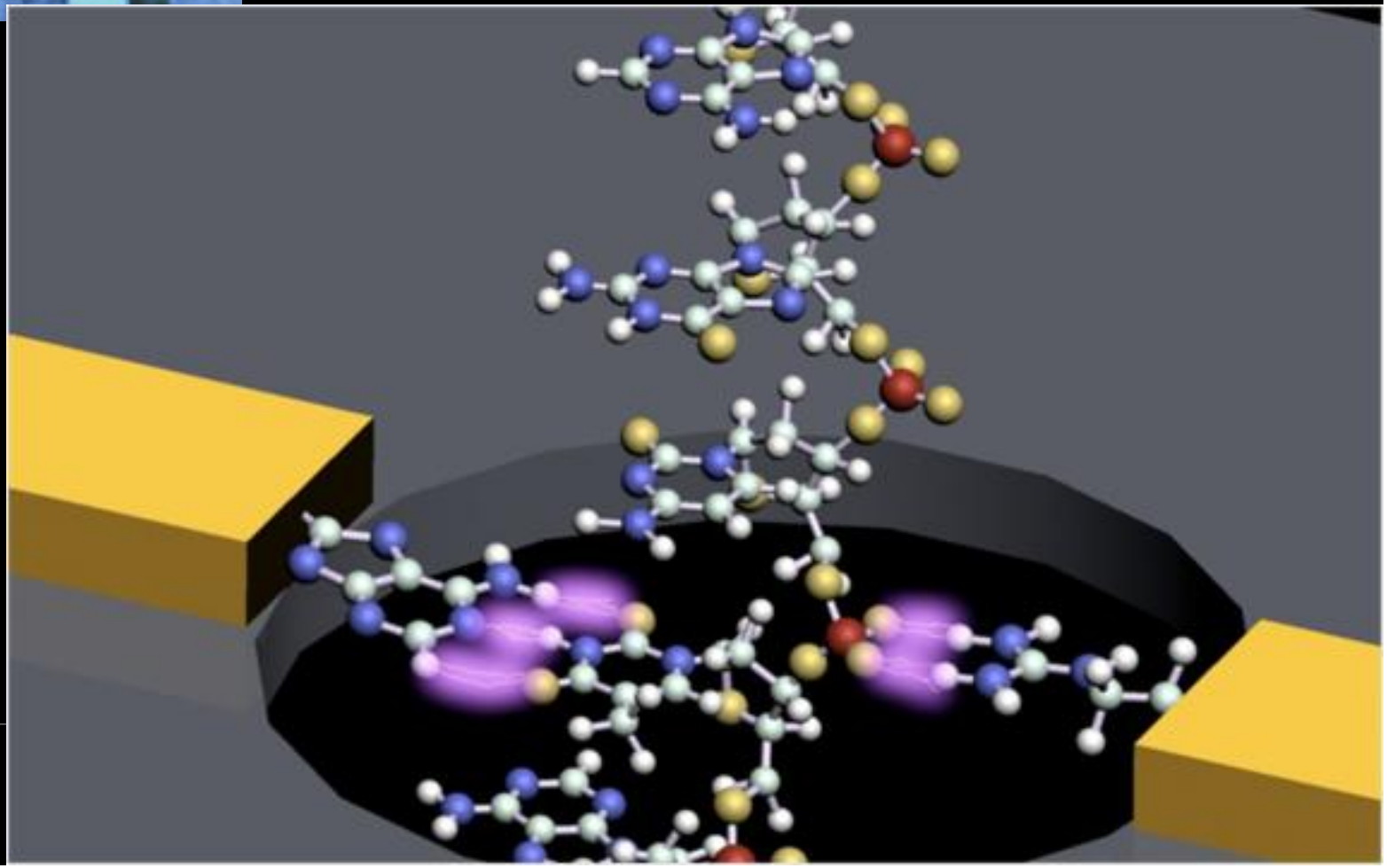
~10% of therapies in Phase I trials will lead to approval

Several specific disease efforts spending ~ $1/3 Billion/year or more to develop therapies


Chemical or
γ Radiation
BAX
BID
Gate Open
Gate Close
BAX
BCL-2 BCL-XL
BCL-2 BCL-XL
Ser136
P
BAD
BAD
APAF-1
CYTOCHROME C
P
PRO-CASPASE 9
AKT
P Thr
P Ser
AKT Pathway
Activation of Effector
CASPASES
CASPASE 3
CASPASE 6
CASPASE 7
Substrate
Cleavage
FAK
PAXILLIN
STAT1
PARP
PKC
TALIN
EIF2α
CASPASES
APOPTOSIS
Modified Signaling
BAX
P
P53
P
ATM


ORIENS
TRIO
SEPTEN
OCCIDENS

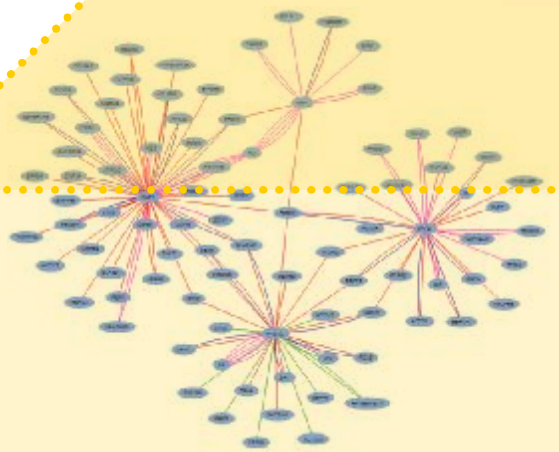

Naval Observatory, Washington, DC



One Dimensional Slices

ENVIRONMENT

Non-coding RNA network

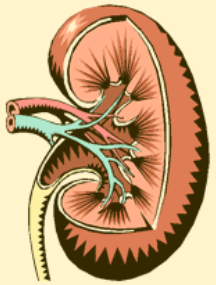


protein network

BRAIN

HEART

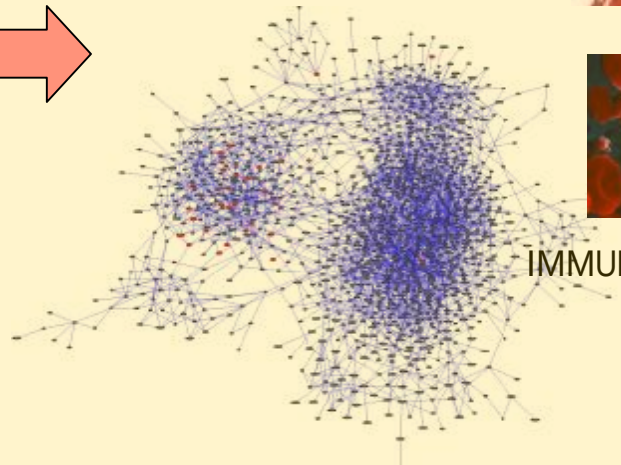
GI TRACT



KIDNEY

ENVIRONMENT

ENVIRONMENT

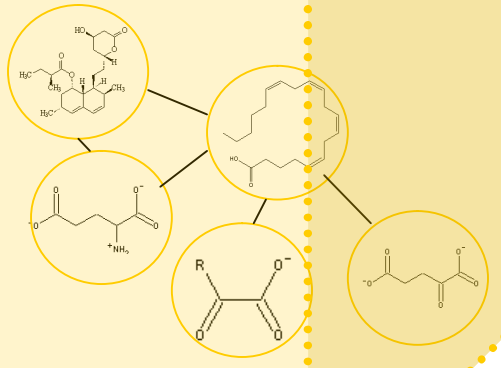


transcriptional network

IMMUNE SYSTEM

VASCULATURE

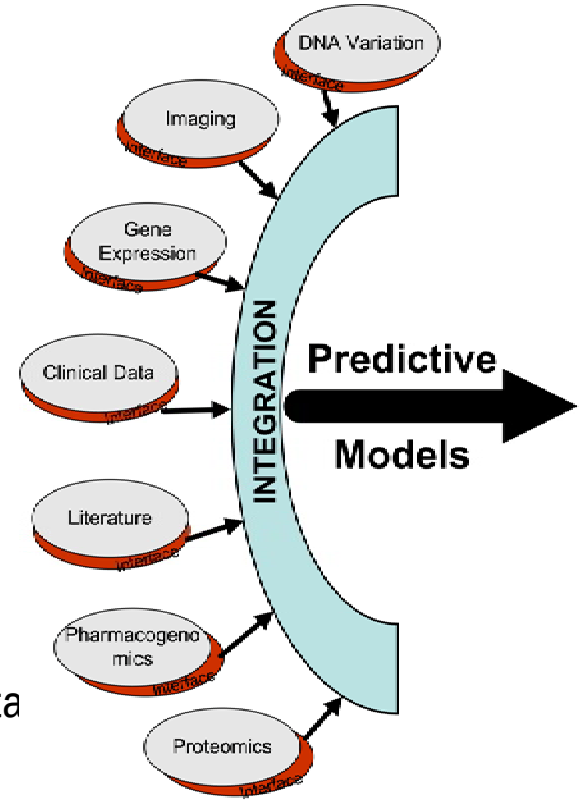
metabolite network

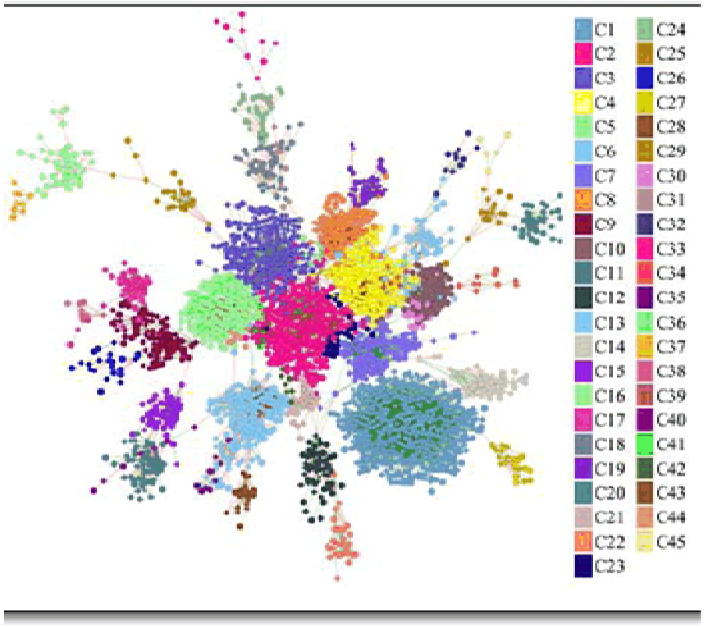


ENVIRONMENT

The “Rosetta Integrative Genomics Experiment”: Generation, assembly, and integration of data to build models that predict clinical outcome

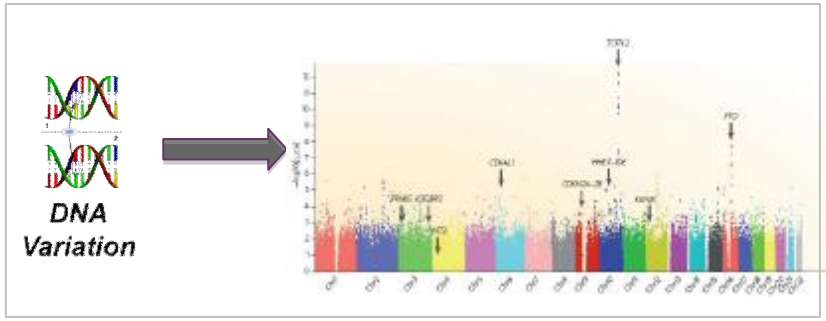
Merck Inc. Co.
5 Year Program
Based at Rosetta
Total Resources
>$150M



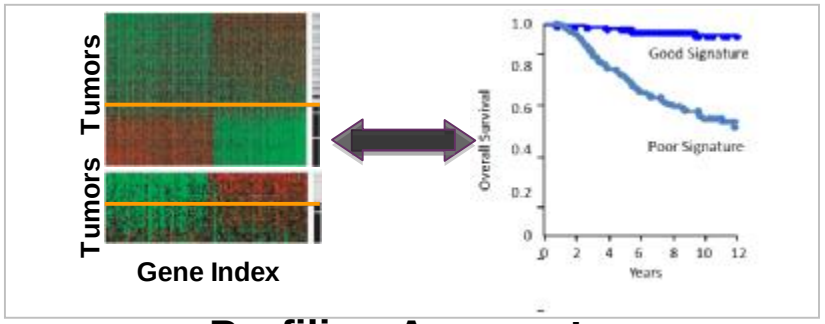


- Generate data need to build
- bionetworks
- Assemble other available data
- Integrate and build models
- Test predictions
- Develop treatments
- Design Predictive Markers

How is genomic data used to understand biology?



"Standard" GWAS Approaches

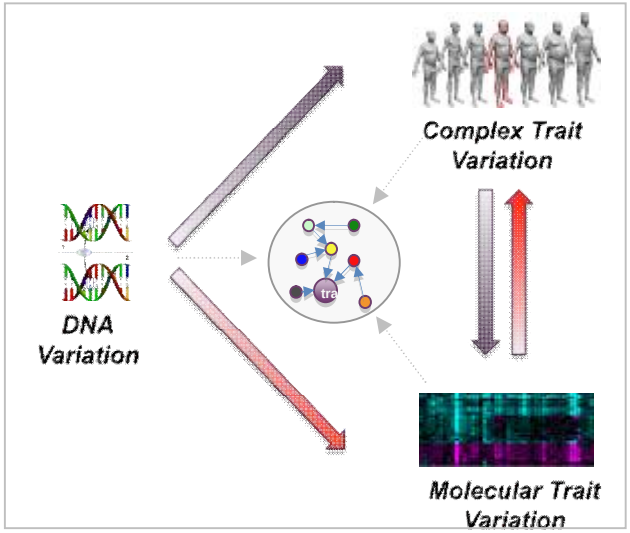


Profiling Approaches

Identifies Causative DNA Variation but provides NO mechanism

Genome scale profiling provide correlates of disease

Ø Many examples BUT what is cause and effect?



"Integrated" Genetics Approaches

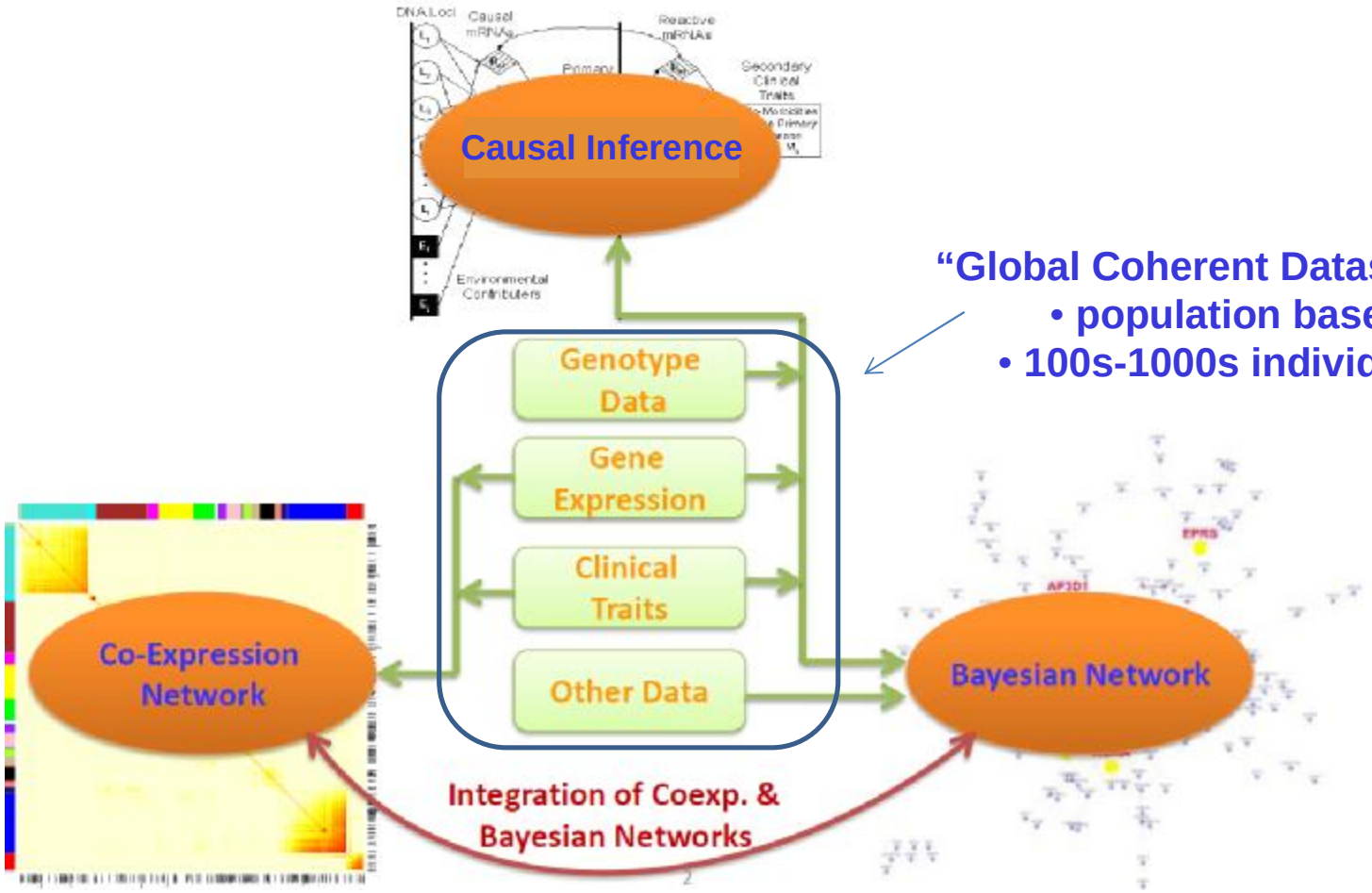
Ø Provide unbiased view of molecular physiology as it relates to disease phenotypes

Ø Insights on mechanism

Ø 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)

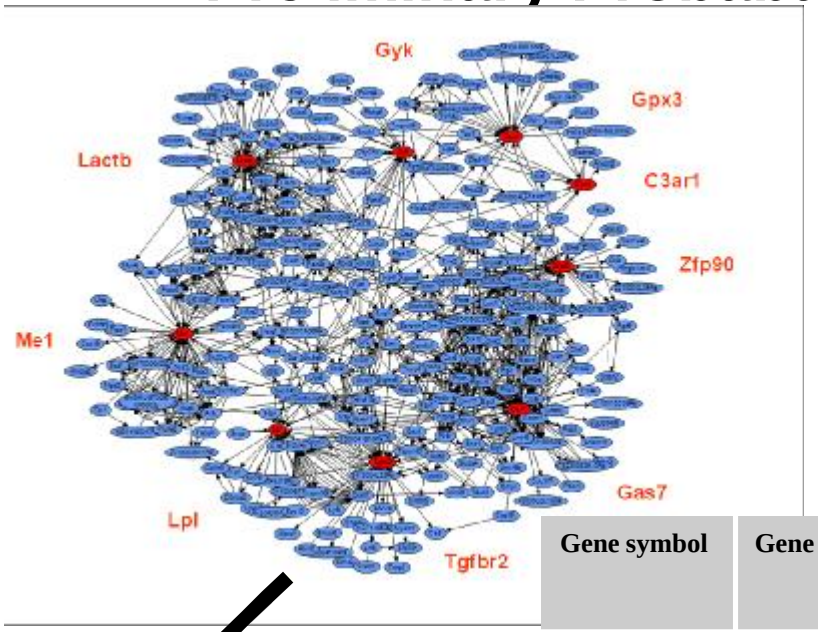


"Global Coherent Datasets"
• population based
• 100s-1000s individuals

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)

Preliminary Probabalistic Models- Rosetta /Schadt



Networks facilitate direct identification of genes that are causal for disease
Evolutionarily tolerated weak spots

Gene symbol	Gene name	Variance of OFPM explained by gene expression*	Mouse model	Source
Zfp90	Zinc finger protein 90	68%	tg	Constructed using BAC transgenics
Gas7	Growth arrest specific 7	68%	tg	Constructed using BAC transgenics
Gpx3	Glutathione peroxidase 3	61%	tg	Provided by Prof. Oleg Mirochnitchenko (University of Medicine and Dentistry at New Jersey, NJ) [12]
Lactb	Lactamase beta	52%	tg	Constructed using BAC transgenics
Me1	Malic enzyme 1	52%	ko	Naturally occurring KO
Gyk	Glycerol kinase	46%	ko	Provided by Dr. Katrina Dipple (UCLA) [13]
Lpl	Lipoprotein lipase	46%	ko	Provided by Dr. Ira Goldberg (Columbia University, NY) [11]
C3ar1	Complement component 3a receptor 1	46%	ko	Purchased from Deltagen, CA
Tgfbr2	Transforming growth factor beta receptor 2	39%	ko	Purchased from Deltagen, CA

Nat Genet (2005) 205:370

Extensive Publications now Substantiating Scientific Approach Probabilistic Causal Bionetwork Models

- **>60 Publications from Rosetta Genetics Group (~30 scientists) over 5 years including high profile papers in PLoS Nature and Nature Genetics**

Metabolic Disease

"*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

CVD

***"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

Bone

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

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

Methods

"*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.**

- #1 - Connect associated SNP to true gene underlying mechanism via Genetics of Gene Expression
  - Workflow - Start with a GWAS or other association between DNA variation and a clinical phenotype, need to understand what genes and ultimately mechanism underlie that association. Here we use our human eSNPs, SNP-set-enrichment, mouse causal genes, and similarities between human and mouse networks to determine plausible genes and network neighborhoods through which the information encoded in that DNA variation manifests as phenotype.

- #2 - Identify new targets and progress through validation as disease genes toward pharmacologic validation
  - Workflow – Predicting genes that contribute to disease phenotypes using causality and network modeling. Multiple examples that validate based on a single-gene intervention in a model system, and ultimately progresses toward in vivo pharmacology.

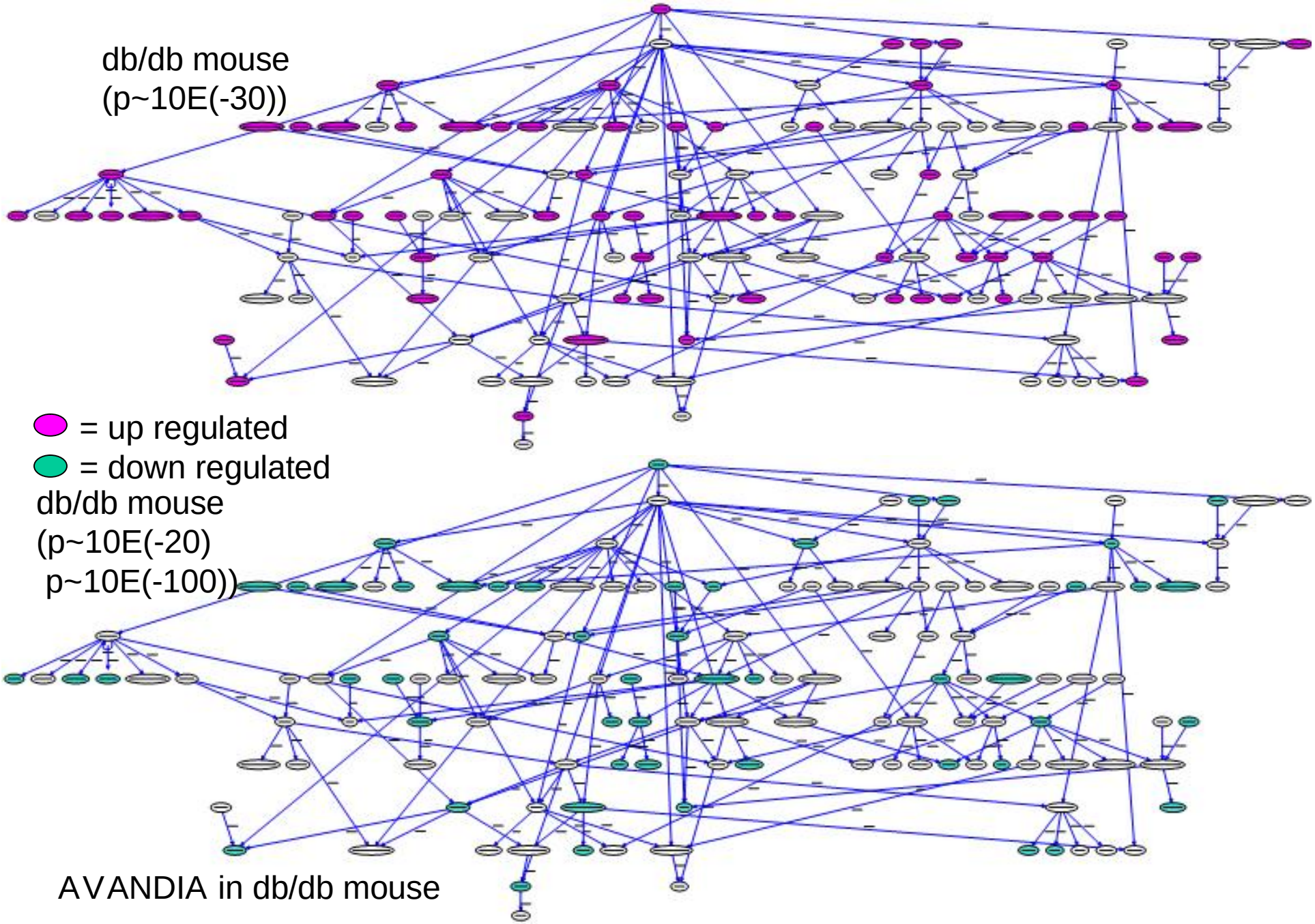
- #3 - Reposition a drug
  - Workflow - Really a special case of the new target identification, where the workflow starts with a number of targets for which good, "safe" compounds exist, and then we apply all the standard approaches we have to validate the target and test the compound for an indication in preclinical species or humans

- #4 - Kill a compound with confidence that opportunities to segment the target population were fully explored.
  - Workflow - Take Phase II or III trial where efficacy is not seeming strong, or where adverse experiences appear mechanism-based. Then use genetics in the trial + the network approaches outlined in case #1 above to demonstrate that a significant segment of the population for which the drug would have substantial net benefit is unlikely to exist.

- #5 - Define clinically relevant subpopulations
  - Workflow - Similar to #4 above, but typically starting at an earlier stage to incorporate hypotheses about population segments early enough in the development process that they are easily tested prospectively.

- #6 - Avoid liability
  - Workflow - Apply a pipeline of standard checks to expression profiling from knockout, siRNA, and compound treatments for a target that encompasses mapping the expression signatures to all relevant tissue networks, looking to see what annotations and other gene expression signatures map to the modules where those intervention signatures map, and following up any leads.

Our ability to integrate compound data into our network analyses



Impact on Merck Pipeline

"The investment has paid off for us."

--Peter Kim, president of Merck Research Laboratory

'The company now has in clinical trials eight drugs that emerged out of Rosetta's platform, Dr. Kim said,with more than a dozen others in preclinical trials. He declined to provide specifics about the costs of the candidate drugs. '

'Dr. Kim said that Merck was developing some cancer drugs that would be directed at various subpopulations of patients rather than the one-size-fits-all approach that has been a hallmark of modern pharmaceutical companies. "We're going to target specific networks and pathways," he said. '

NY Times, August 25, 2009

details at:
http://sagebase.org/research/publications.html

Assembling Networks for Use in the Clinic

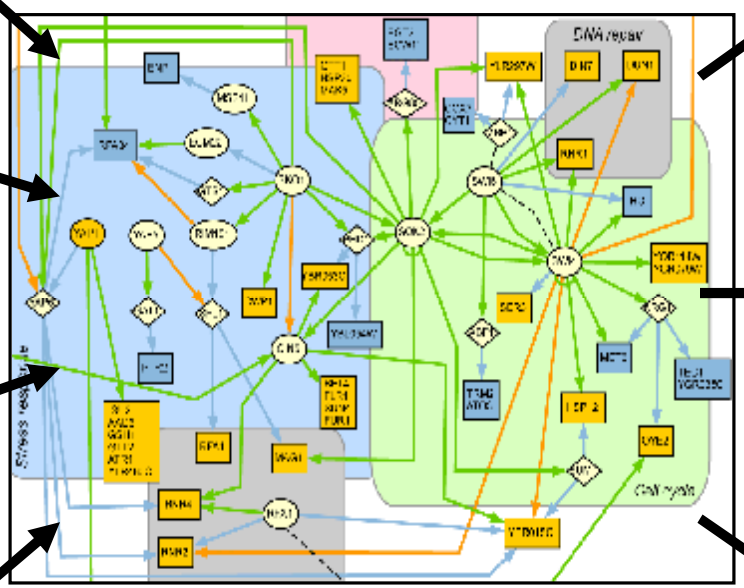
Network evolutionary comparison / cross-species alignment to identify conserved modules

Projection of molecular profiles on protein networks to reveal active modules

Integration of transcriptional interactions with causal or functional links

Alignment of physical and genetic networks

The Working Map



Network-based cancer diagnosis / prognosis

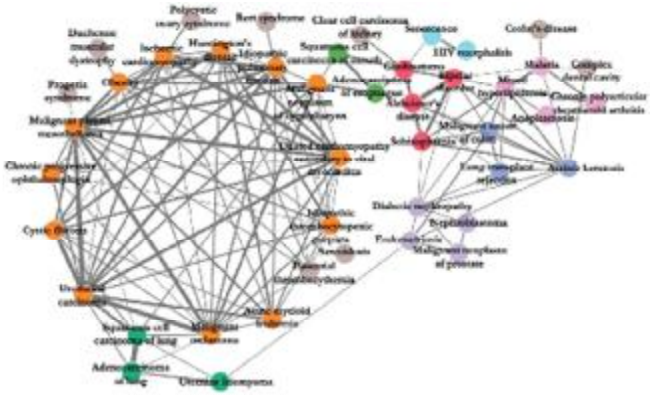
Identification of networks associated with cancer progression

Moving from genome-wide association studies (GWAS) to network-wide "pathway" association (NWAS)

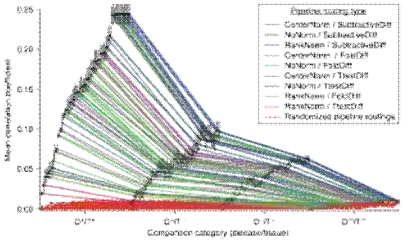
Pathway assembly via integration of networks → **Network based study of disease**

Exploring the Global Landscape of Human Disease Through Public Data

Public data enables quantitative disease relationships



Silpa Suthram et al. PLoS computational biology (2010) vol. 6 (2) pp. e1000662

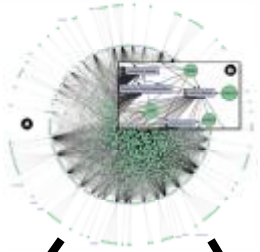


Joel Dudley et al.. Molecular systems biology (2009) vol. 5 pp. 307

High quality signals exist in public data

Differences ⟷ Commonalities

Plasma proteome networks

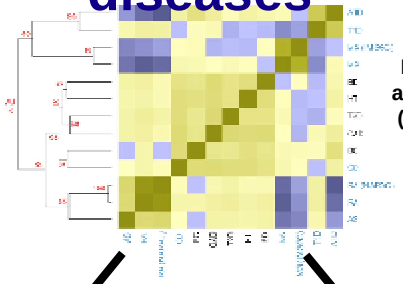


Joel Dudley and Atul Butte. Pacific Symposium on Biocomputing (2009) pp. 27-38

Which biomarkers best discriminate diseases?

Is there a blood biomarker for general pathology?

Genetic architecture of autoimmune diseases

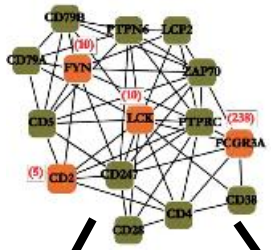


Marina Sirota et al. PLoS genetics (2009) vol. 5 (12) pp. e1000792

Are there genetic "switches" for autoimmunity ?

Is there a common autoimmune susceptibility variant?

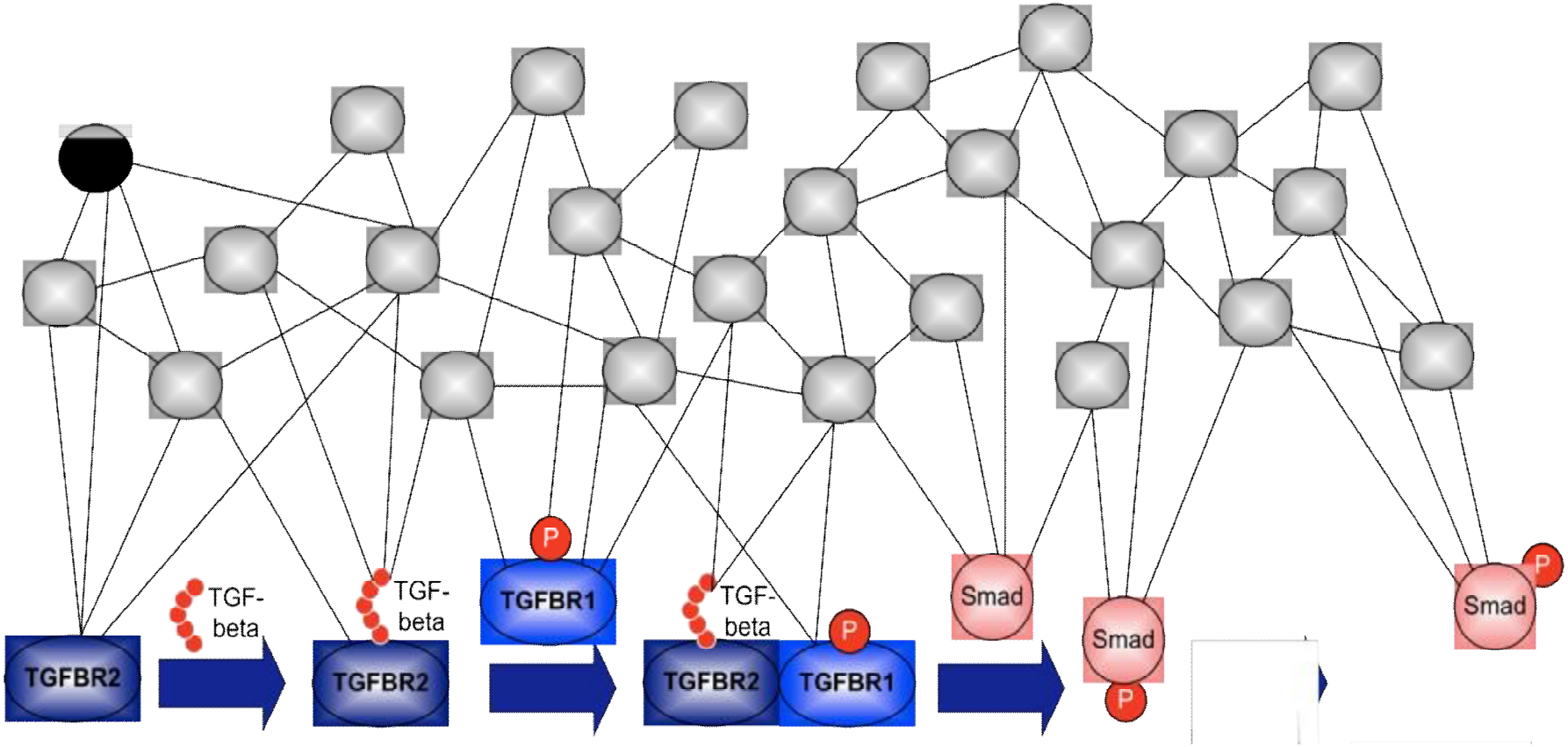
Functional gene module networks



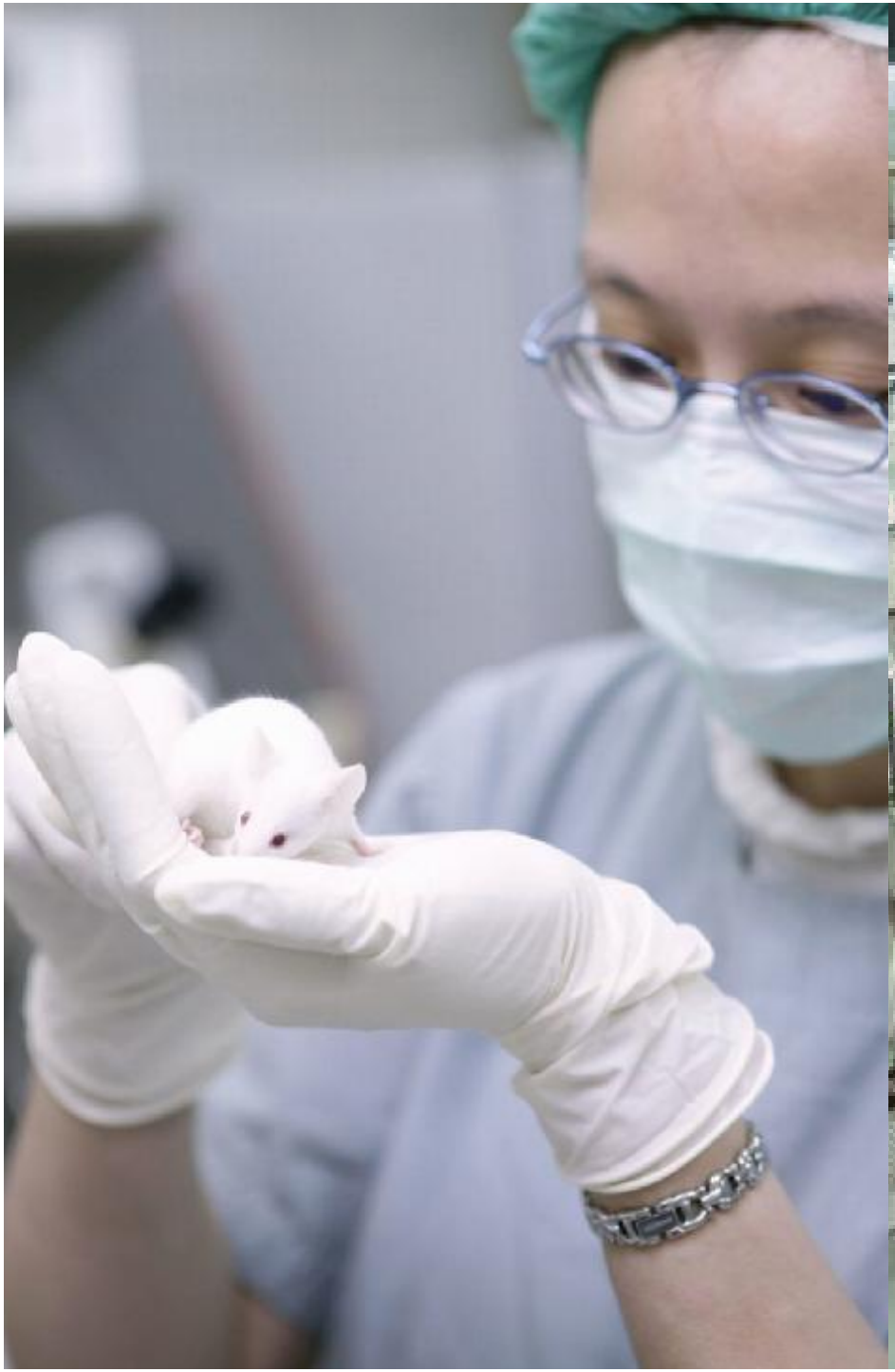
Silpa Suthram et al. PLoS computational biology (2010) vol. 6 (2) pp. e1000662

Which modules are unique to metabolic diseases?

Do common modules harbor pluripotent drug targets?


TGF-
beta
TGFBR2
TGF-
beta
TGFBR2
P
TGFBR1
TGF-
beta
TGFBR2
P
TGFBR1
Smad
Smad
P
Smad
P

what we see...



The stunning technologies coming will generate heaps of genomic data poised to

Bionetworks using integrative genomic approaches can highlight the non-redundant components- can find drivers of the disease and of therapies

Need to develop ways to host massive amounts of data, tools, evolving representations of disease as represented by these probabilistic causal disease models

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

Willingness at Merck to imagine a world where all of **disease biology was considered precompetitive space**.

Realizing the donation by Merck **might seed a "commons"** allowing a potential long term gain to the whole community provided by evolving models of disease built via a contributor network

Sage Mission

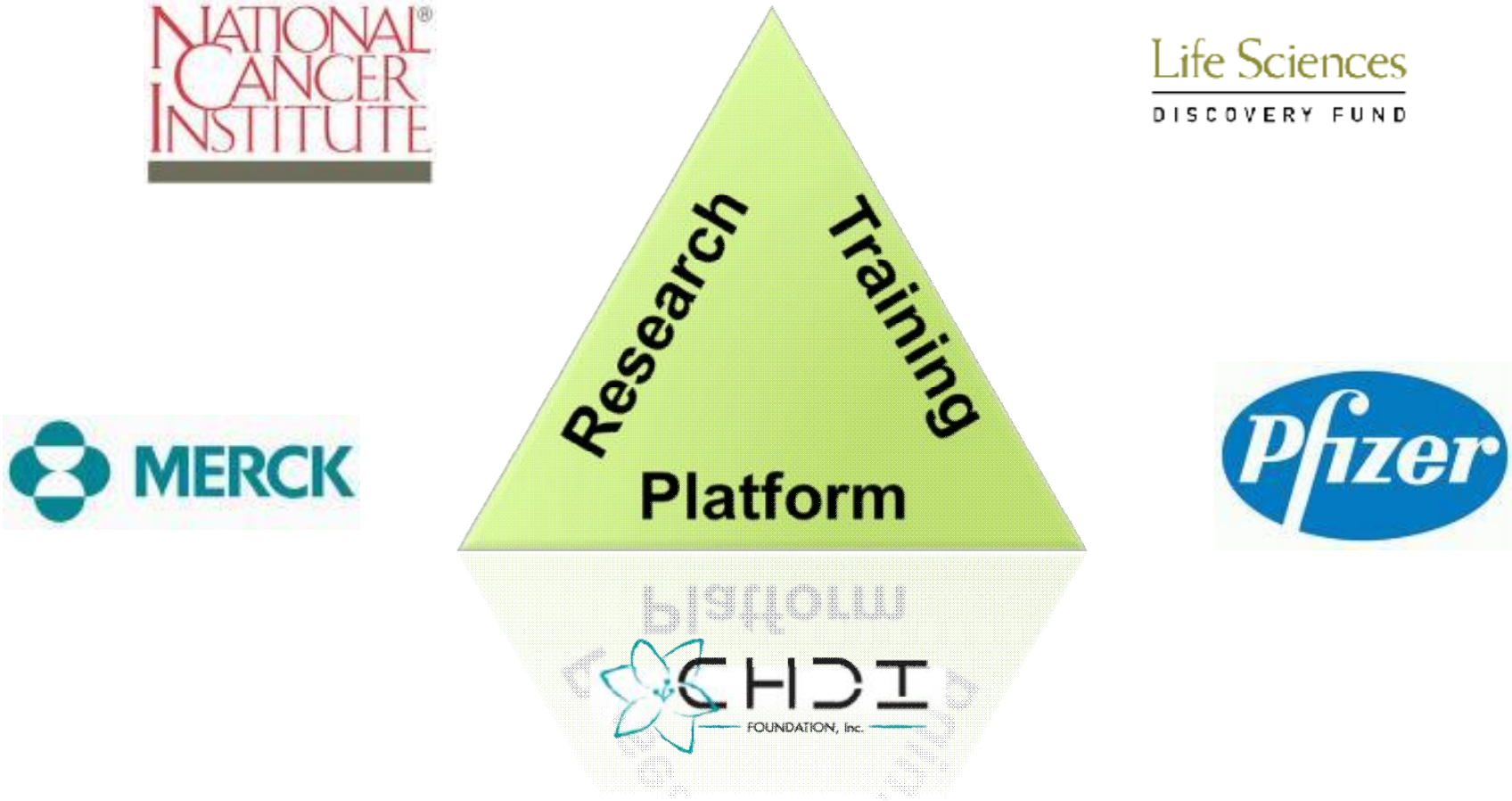
Sage Bionetworks is a non-profit organization with a vision to create a “commons” where integrative bionetworks are evolved by contributor scientists with a shared vision to accelerate the elimination of human disease

Sage Bionetworks

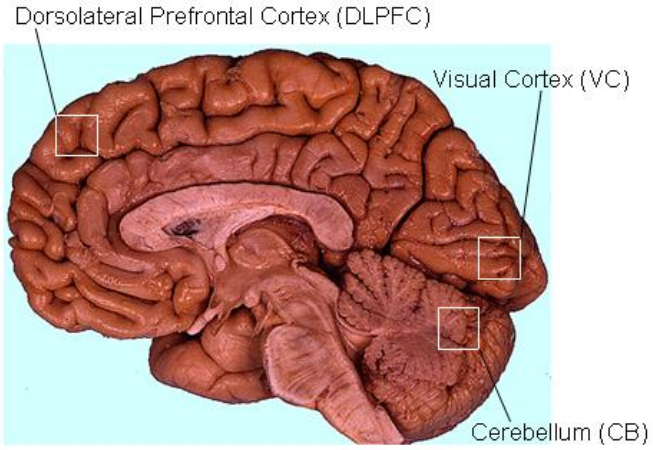
FRED HUTCHINSON
CANCER RESEARCH CENTER

A LIFE OF SCIENCE

Sage Bionetworks



Example 1: Alzheimer's Disease

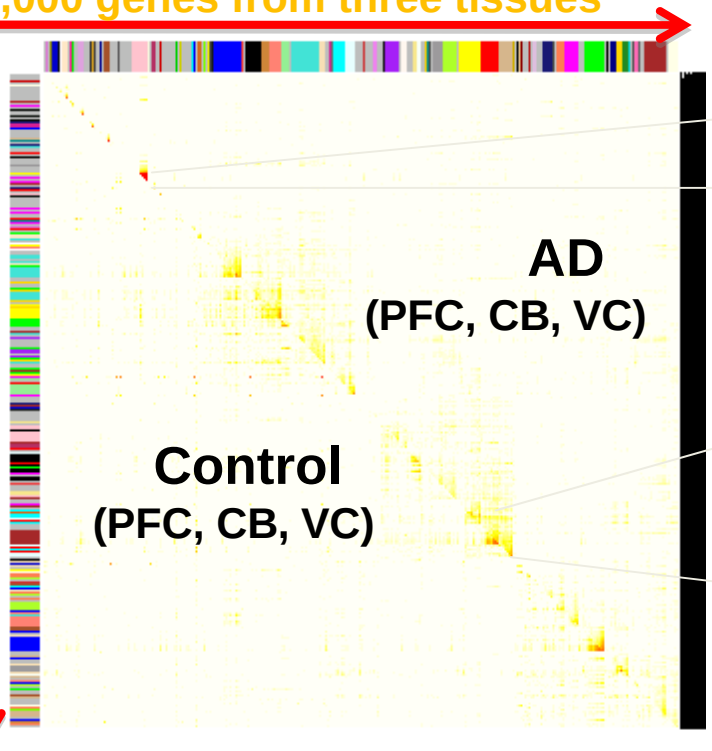


- Ø Cross-tissue coexpression networks for both normal and AD brains
  - prefrontal cortex, cerebellum, visual cortex
- Ø Differential network analysis on AD and normal networks
- Ø Integrate coexpression networks and Bayesian networks to identify key regulators for the modules associated with AD

subset	samples
Alzh_PFC	310
Alzh_CR	263
Alzh_VC	190
Norm_PFC	153
Norm_CR	128
Norm_VC	121

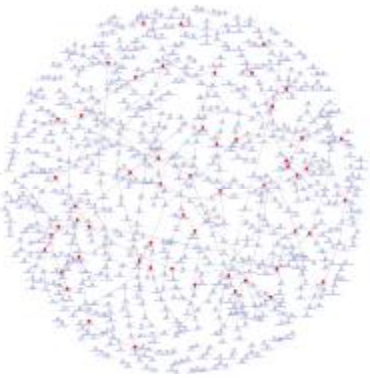
Identification of Disease (AD) Pathways via Comparative Gene Network Analysis

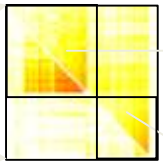
40,000 genes from three tissues



Glutathione transfer

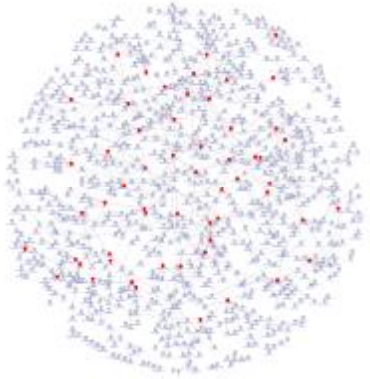
Gain connectivity by 91 fold





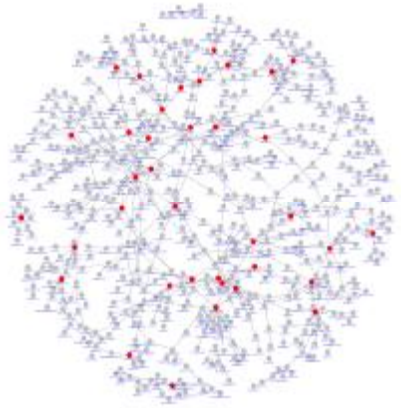
nerve ensheathment

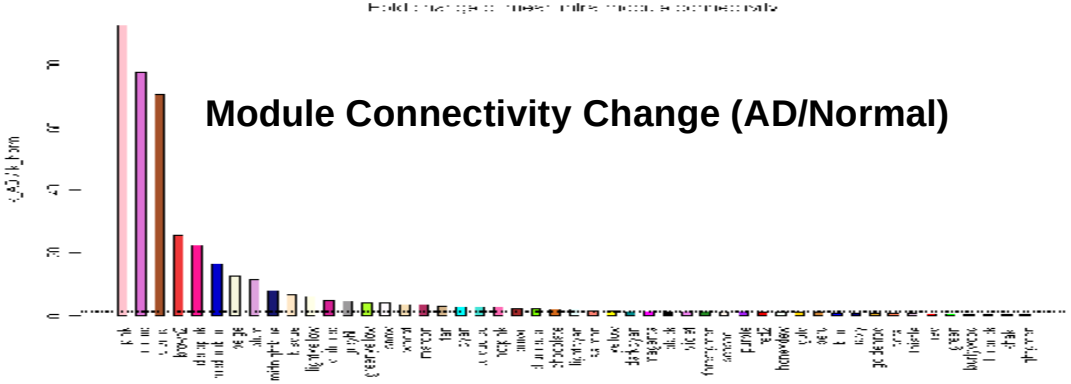
Lose connectivity by 40%



extracellular matrix

Gain connectivity by 1.9 fold





Module Connectivity Change (AD/Normal)

Bayesian Subnetworks

Key Regulators

GlutathioneTransferase NerveEnsheathment ExtracellularMatrix

pink	hits	red	hits	tan	hits
PECAM1.VC	70	ENPP2.PFC	296	SLC22A2.PFC	238
XM_211501.VC	62	PLLP.PFC	135	OGN.PFC	120
GON4L.VC	52	PLP1.PFC	133	KIAA1199.PFC	83
GNPTAB.VC	45	FRYL.PFC	129	AK021858.PFC	77
GSTA4.VC	45	SLC44A1.PFC	129	Contig39710_RC.PFC	66
hCT24928.VC	41	Contig43380_RC.PFC	125	SPTLC2L.PFC	64
RAB2.VC	41	PLEKHH1.PFC	123	COL6A3.PFC	62
HIST1H2BA.VC	38	UGT8.PFC	118	PTGDR.PFC	54
ENST00000283038.VC	35	AL137342.PFC	112	XM_068880.PFC	48
hCT1959721.VC	35	TTYH2.PFC	87	NM_018242.PFC	47
OR6S1.VC	31	PSEN1.PFC	73	SVIL.PFC	47
DOCK6.VC	30	TRIM59.PFC	73	CLIC6.PFC	43
ENST00000293571.VC	28	FA2H.PFC	69	OLFML2A.PFC	31
OR12D3.VC	28	KIAA1189.PFC	61	MYH11.PFC	27
AK055724.VC	27	CREB5.PFC	59	MRC2.PFC	26
Contig33276_RC.VC	25	AB037815.PFC	57	Contig16712_RC.PFC	25
hCT1658538.VC	25	MAP7.PFC	46	WNT6.PFC	25
ABCC2.VC	23	ABCA2.PFC	41	C1S.PFC	21
AK057434.VC	19	NM_014711.PFC	41	DAB2.PFC	20
hCT1660876.VC	17	NM_175922.PFC	39	PCOLCE.PFC	20
MYOHD1.VC	17	FRMD4B.PFC	38	SLPI.PFC	19
hCT1644335.VC	16	RTKN.PFC	36	Contig47865.PFC	17
HSS00083045.VC	16	NM_144595.PFC	35	FCGR2B.PFC	15
PIGV.VC	16	FOLH1.PFC	34	TBX15.PFC	14
RAC3.PFC	16	SEPT4.PFC	32	COL3A1.PFC	12
WDR23.PFC	16	LAMP2.PFC	31	SCARA5.PFC	12

PECAM1: Platelet-endothelial cell adhesion molecule, a tyrosine phosphatase activator that plays a role in the platelet activation, increased expression correlates with MS, Crohn disease, chronic B-cell leukemia, rheumatoid arthritis, and ulcerative colitis

ENPP2: Phosphodiesterase I alpha, a lysophospholipase that acts in chemotaxis, phosphatidic acid biosynthesis, regulates apoptosis and PKB signaling; aberrant expression is associated with Alzheimer type dementia, major depressive disorder, and various cancers

SLC22A25: solute carrier family 22, member 25, Protein with high similarity to mouse Slc22a19, which is a renal steroid sulfate transporter that plays a role in the uptake of estrone sulfate, member of the sugar (and other) transporter family and the major facilitator superfamily

Glutathione Transferase Module (Pink)

- 983 probes from all three brain regions (9% from CB, 15% from PFC and 76% from VC)
- Most predictive of Braak severity score

Example 2: Sage Non-Responders Cancer Project

Patient Oriented Cohort Study to ID Non-Responders to Approved CA Drugs

Co-Chairs- Stephen Friend Charles Saywers Todd Golub & Rich Schilsky

Multiple Myeloma- Ken Anderson/ DFCI- Kathy Guiste/MMRF
AML at First Relapse- Fred Applebaum /FHCRC- Louis deGennaro/LLS
Non-Small Cell Lung Cancer- Roy Herbst MD Anderson / “LCA”
Ovarian Cancer- Beth Karlan/Cedar Sinai- Laura Shawver:/Clearity Foundation
Breast Cancer- Laura Esserman/UCSF- “TBD”

Molecular Profiling: Levine, Polit, Levine

Patient Outreach: Live Strong-Lance Armstrong Foundation/ 23andMe

Platform: Global Coherent Datasets

A data set containing genome-wide DNA variation and intermediate trait, as well as physiological phenotype data across a population of individuals large enough to power association or linkage studies, typically 50 or more individuals. To be coherent, the data needs to be matched with consistent identifiers. Intermediate traits are typically gene expression, but may also include proteomic, metabolomic, and other molecular data.

Status	Definition
Sage - Available	Dataset available from Sage website
Sage - Transition	Dataset in process of being made available
Requires Release	Dataset with known or anticipated legal release requirements prior to posting on Sage website
In progress	Dataset not yet complete

GCDs are current state of knowledge and subject to change as more information becomes available to Sage

Platform: Models available in Sage Repository

Dataset	Clinical	Genotype	Expression	Copy Number Variations	Networks
Human Cancer Breast BCCA	No	No	No	No	Bayesian and Coexpression
Mouse CVD Adipose, Liver , Brain, Muscle UCLA	Yes	Yes	Yes	No	Bayesian and Coexpression
Human CVD Liver Vanderbilt/ Pittsburg/St Judes	Yes	dbGaP	Yes	No	Bayesian and Coexpression
Differentiating ES cell regulation	No	No	No	No	Interaction
Human B-Cell Interactome	No	No	No	No	Interaction
Human Cancer HCC HKU	Yes	No	Yes	No	Bayesian and Coexpression
Human Cancer Glioblastoma TCGA	No	No	No	No	Bayesian and Coexpression
Yeast Genetic Interaction Map	No	No	No	No	Interaction

Platform: Tools- Download Page for Repository

Sage BIONETWORKS

Home | Company | Research | Commons | Training

Home › Research › Tools

Key Driver Analysis Tool

Overview:

Key Driver Analysis (KDA) is an analysis tool, as both an R package and Cytoscape plugin, for identifying key regulators of a gene regulatory network. It takes as input a gene network **N** (directed or undirected) and a gene set (module) **G**. The gene set is any subset of genes from the network **N** (e.g. pathway, module, ontology), permitting focus on a particular biological context.

Prerequisites:

- Java, 5.0+ (www.javasoft.com)
- Cytoscape, 2.6+ (www.cytoscape.org)

Download KDA:
https://sourceforge.net/projects/sagebionetworks/files

Installation:
The KDA archive contains the plugin source, the plugin jar file, and several example datasets. To install, copy the file 'kda-plugin.jar' into Cytoscape's plugin directory [cytoscape_install_path/plugins]. If Cytoscape is open, close and restart cytoscape; the plugin should now be visible from within the "Plugin" menu.

Running KDA (by example):
This tutorial assumes some elementary knowledge of Cytoscape, such as importing and laying out networks. Please read the Cytoscape documentation prior to using KDA.

1. KDA requires an active network within the Cytoscape desktop. The KDA package contains 2 sample networks: open the network labeled "yeastbn.cys" found in the "kda/data" directory [Figure 1].

2. Activate the KDA plugin by selecting the "Key Driver Analysis" menu item from the "Plugin" menu [Figure 2]. This will open a "Key Driver Analysis Settings" dialog [Figure 3]. The dialog contains the following settings:

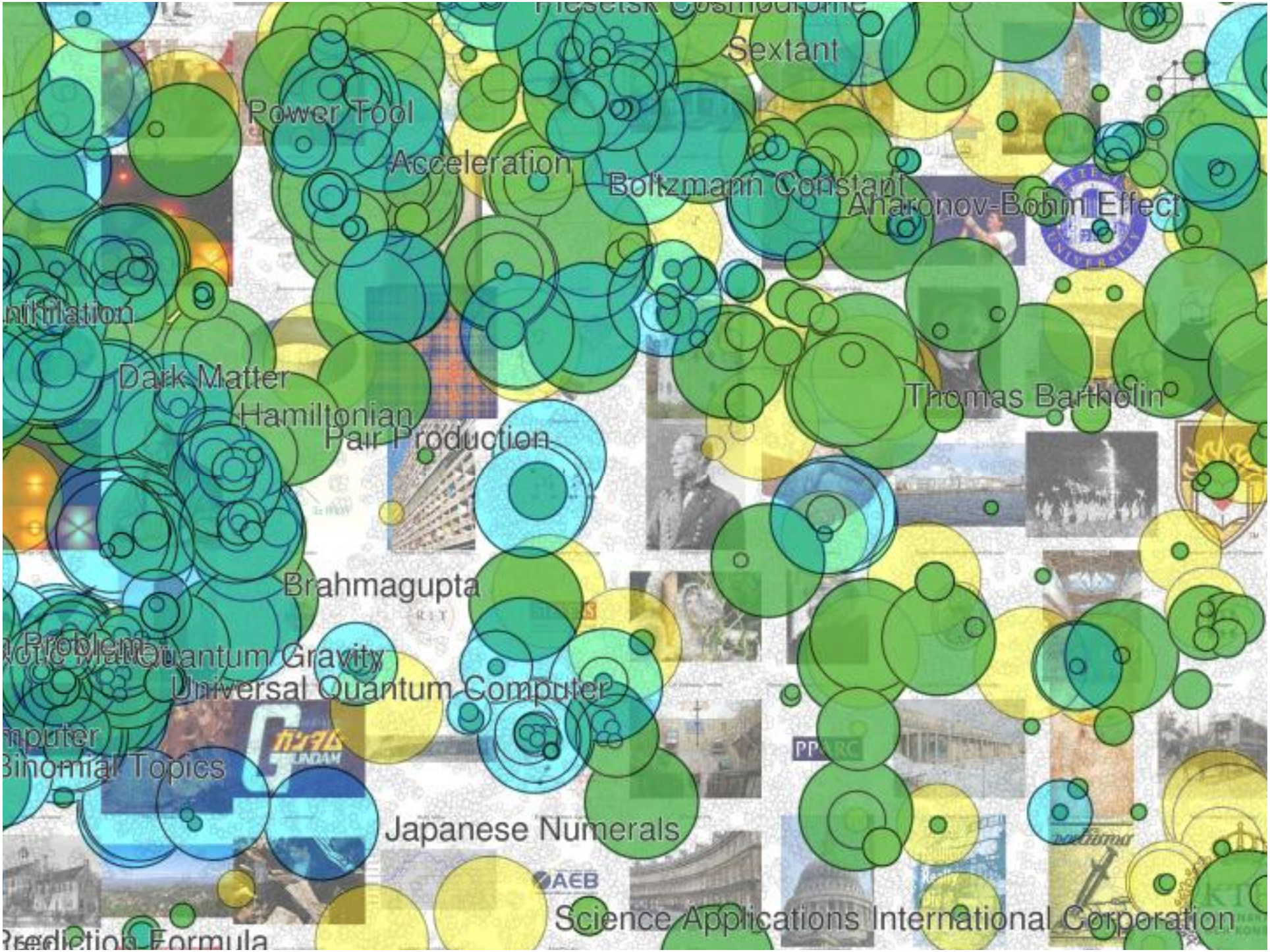
http://www.sagebase.org/research/tools.html

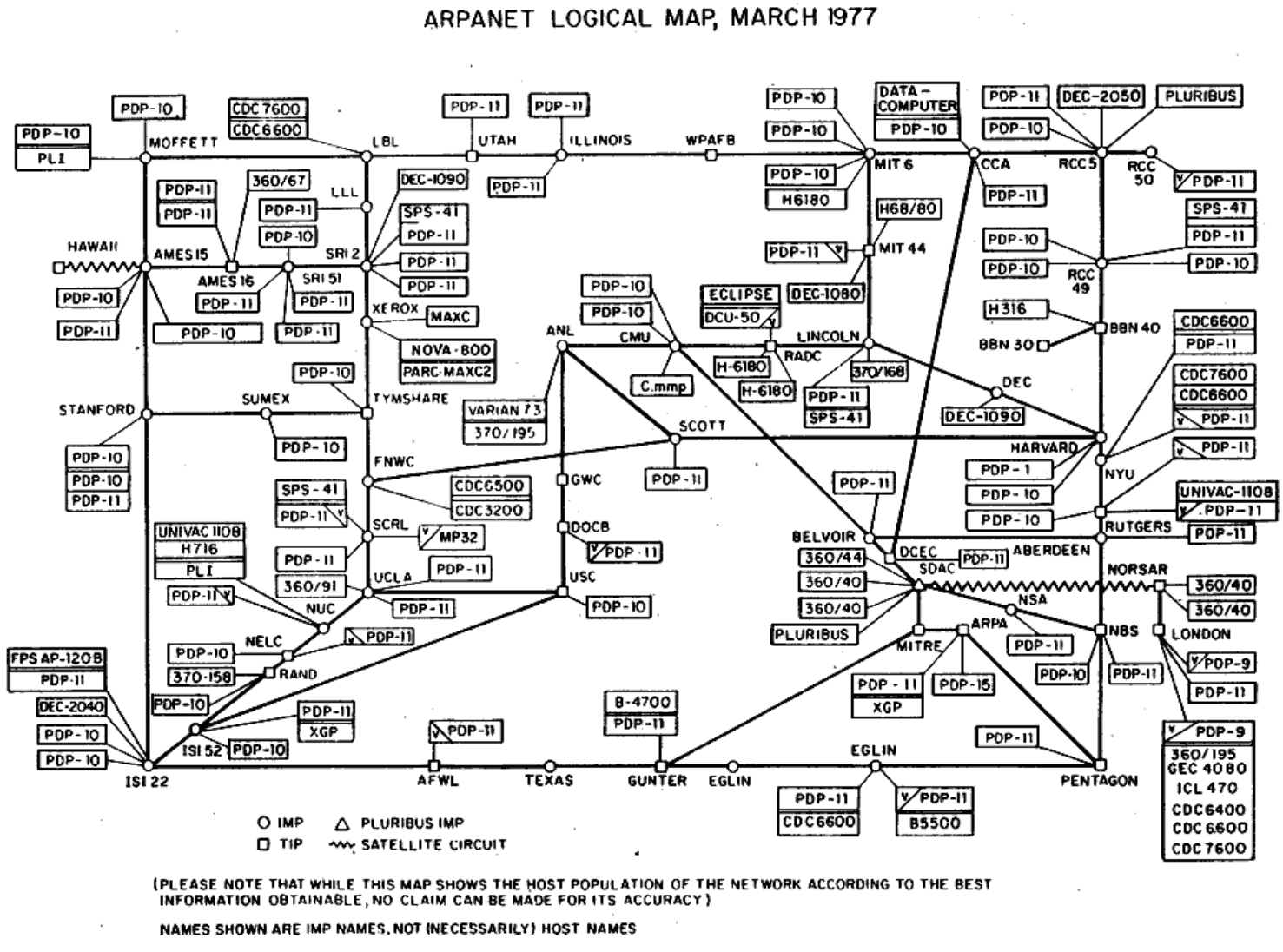
NOT JUST WHAT BUT HOW

data mining
“my data’s mine, and your data’s mine”

attribution: carole goble- sidney brenner


this must be integrated.

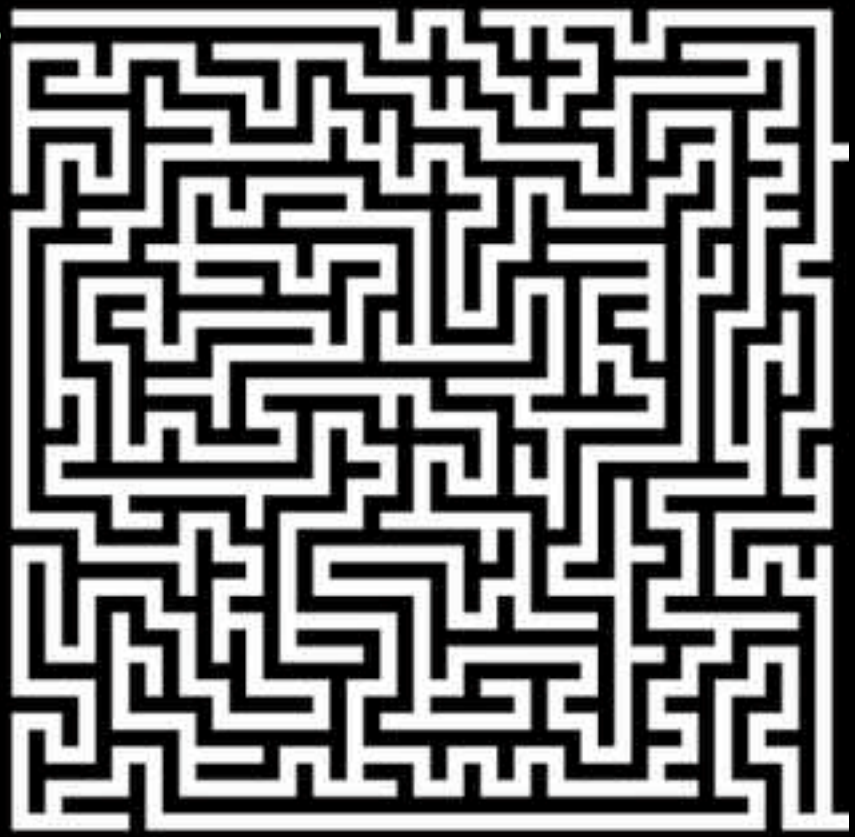

Sextant
Power Tool
Acceleration
Boltzmann Constant
Aharonov-Bohm Effect
Dark Matter
Thomas Bartholin
Hamiltonian
Pair Production
Brahmagupta
Quantum Gravity
Universal Quantum Computer
Binomial Topics
Japanese Numerals
Science Applications International Corporation



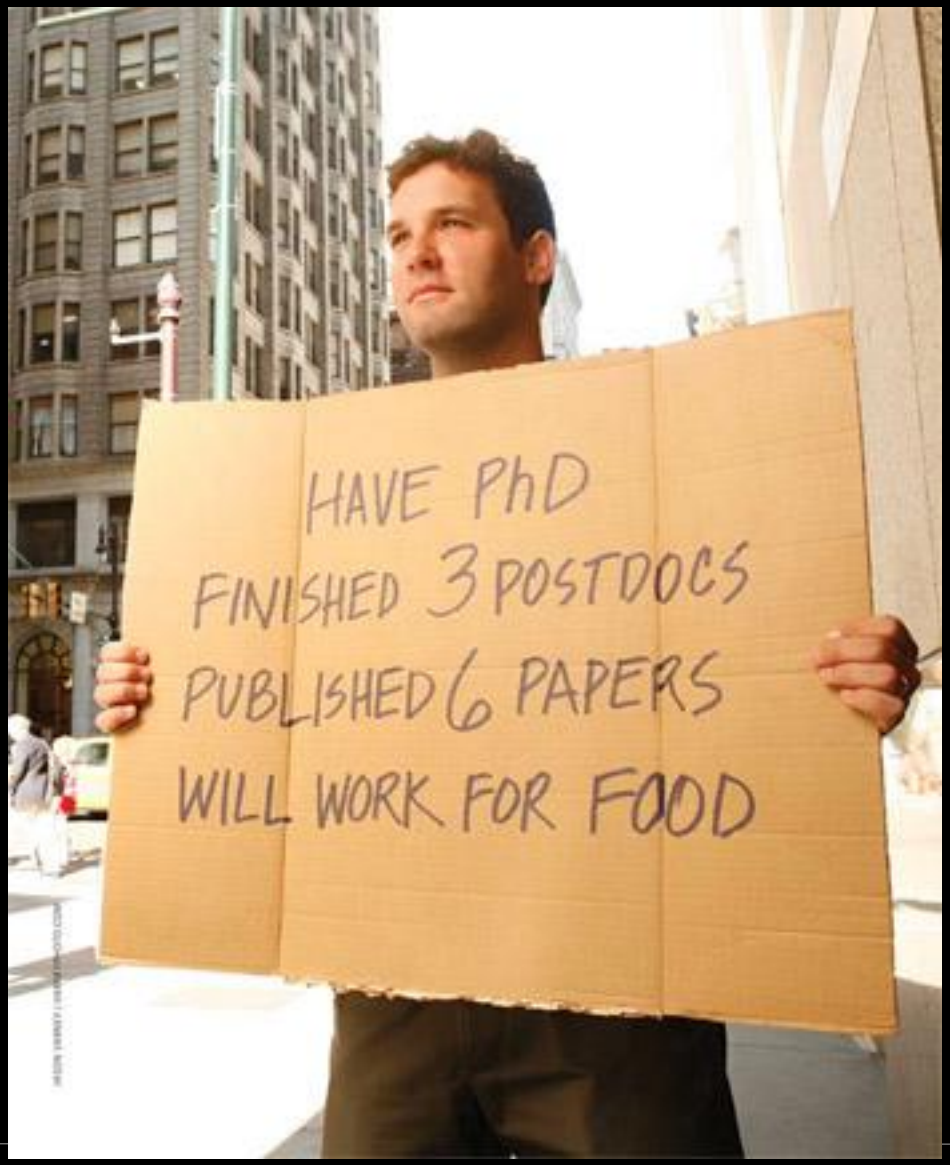
sharing as an adoption of common standards..

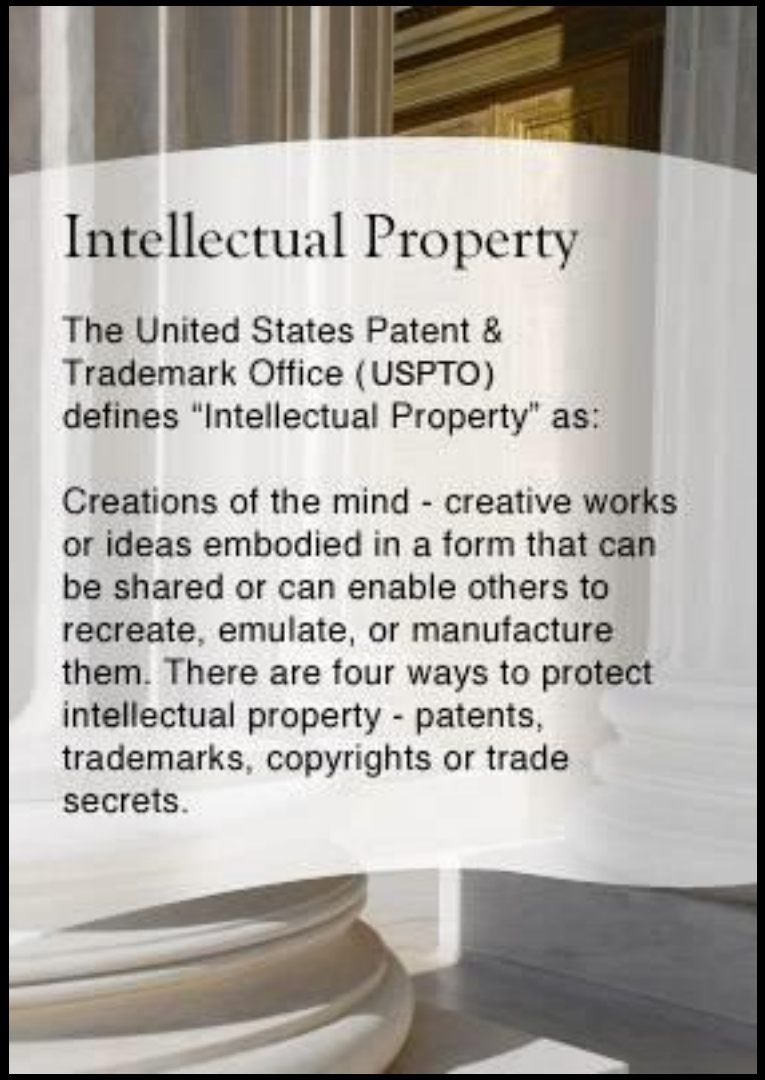
Clinical Genomics Privacy IP

PATIENTS
DATA AND SAMPLES



THERAPIES


HAVE PhD
FINISHED 3 POSTDOCS
PUBLISHED 6 PAPERS
WILL WORK FOR FOOD


Intellectual Property
The United States Patent & Trademark Office (USPTO) defines "Intellectual Property" as:
Creations of the mind - creative works or ideas embodied in a form that can be shared or can enable others to recreate, emulate, or manufacture them. There are four ways to protect intellectual property - patents, trademarks, copyrights or trade secrets.

The Sage Commons

Developing and sharing large scale predictive network models of disease.

The Sage Commons will be a revolutionary accessible information platform to define the molecular basis of disease and guide the development of effective human therapeutics and diagnostics.

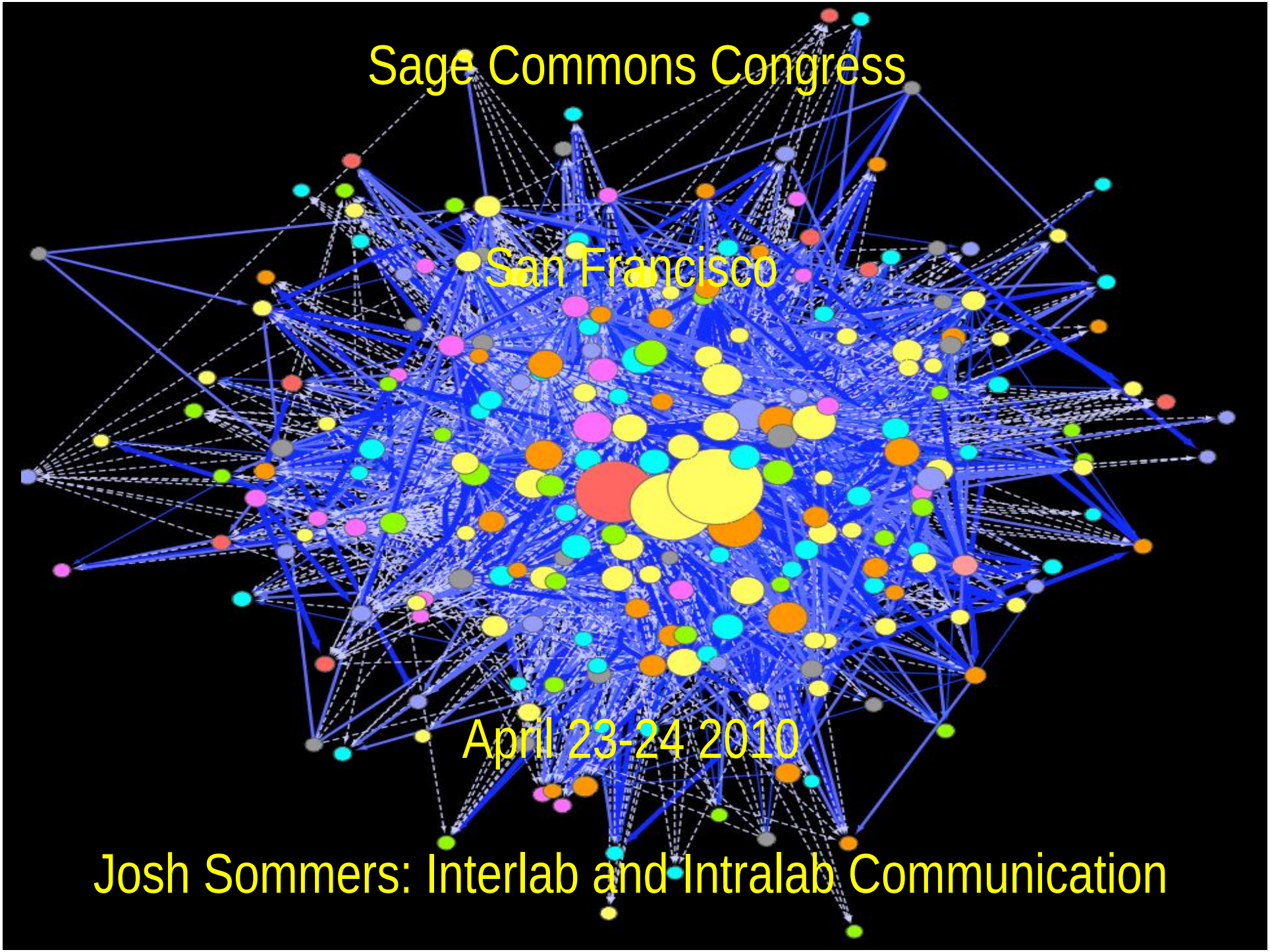
Integration of molecular mega-data sets

The Sage Commons will be used to integrate diverse molecular mega-data sets, to build predictive bionetworks and to offer advanced tools proven to provide unique new insights into human disease biology. Users will also be contributors that advance the knowledge base and tools through their cumulative participation.

The public access goal of the Sage Commons requires the development of a new strategic and legal framework to protect the rights of contributors while providing widespread access to fundamentally non-commercial assets.

Linking human disease biology models

Sage seeks to work with the academic and commercial research communities to satisfy a substantial unmet need in useful human disease biology models. Human disease biology is defined in this context as an understanding of the


Sage Commons Congress
San Francisco
April 23-24 2010
Josh Sommers: Interlab and Intralab Communication

EXTENDING STANDARD AGREEMENTS FOR DATA SHARING-FUNDERS AND PUBLISHERS

All data supporting the publication shall be made available for download from a digital repository under terms and conditions no more restrictive than the Science Commons Protocol for Implementing Open Access Data {http://sciencecommons.org/projects/publishing/open-access-data-protocol/},

upon:

a) six (6) months after any publication describing the results of the funded research project;

b) twelve (12) months after the completion of the research project; or

c) twelve (12) months after the expiration or termination of the Grant

Agreement,whichever is earliest, and subject to any reasonable delay necessary to evaluate for patentability and to file any patent applications. Grantee may comply with the above requirement either by: Depositing a copy of the data in a third party digital repository from which it may be downloaded free of charge, or Offer such data for download on a Website without charge, or Offer to distribute such data on any medium which is commonly used, subject to a reasonable charge for the cost of reproduction and distribution. Deposit of Unpublished Data

Grantee shall deposit a copy of all data created in the course of the funded research project in Grantor's data repository no later than six (6) months from the date of creation. The data so deposited shall be used by Grantor only for its own internal quality analysis and shall not be published by Grantor, until such data otherwise becomes publicly available.

How to Host Network Models

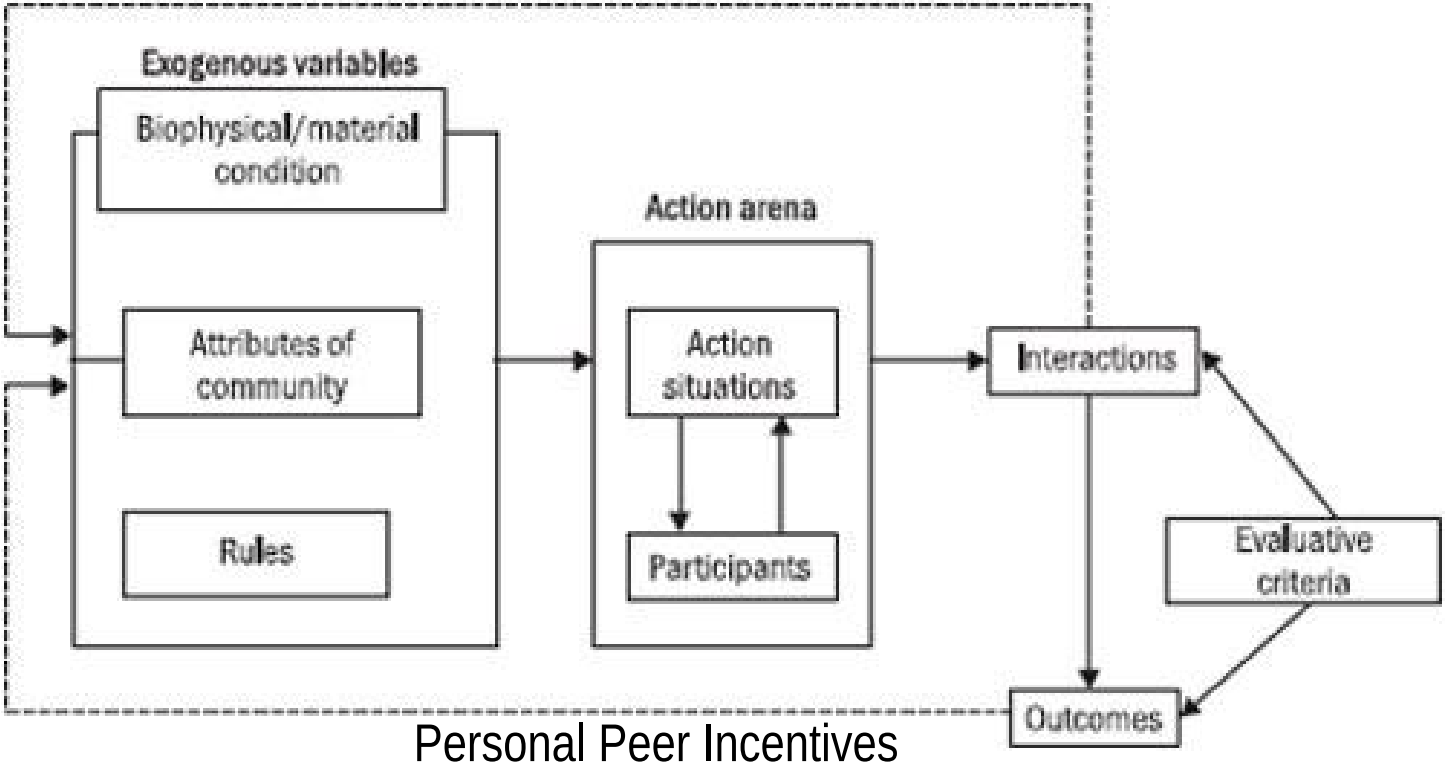
Sage Bionetworks is working on a major agreement with a major Publisher

We will need to develop ways to host massive amounts of data, evolving representations of disease as represented by these probabilistic causal disease models

We will need to learn how to share data, and models and fundamental change how we fund and reward science- head towards a more contributor distributed world

The patient and their disease foundations will be at the center of this world where disease biology will exist in pre-competitive space surrounded by IT partners, knowledge experts NIH, pharma, insurers, diagnostic companies

Institutional Analysis and Development Framework (Ostrom) to Examine the Current Rules and Structures that define Current Approaches to Drug Discovery-could enable the Next Generation Discovery Engines



Personal Peer Incentives

Academic Institutions and their Reward structures

Institutional IP Rules

Pharmaceutical Industry Rewards

Publishers

Venture Capitalists

Physician Trialists

BETTER MAPS OF DISEASE

NOT JUST WHAT BUT HOW

BUILDING A COMMONS FOR EVOLVING GENERATIVE MODELS OF DISEASE

Specific Proposals:

1- Adopt Sharing rules for making data accessible

2- Engage/Support Non-Responders to Approved Oncology Drug Cohort Study

3- Collect Control Arms for Clinical Trials where genomics data (SNPs and RNA expression) was obtained

4- Build Disease specific IP free zone: shared data/tools/models