BETTER MAPS OF DISEASE

NOT JUST WHAT BUT HOW

BUILDING A COMMONS FOR EVOLVING GENERATIVE MODELS OF DISEASE

Existing approaches and issues

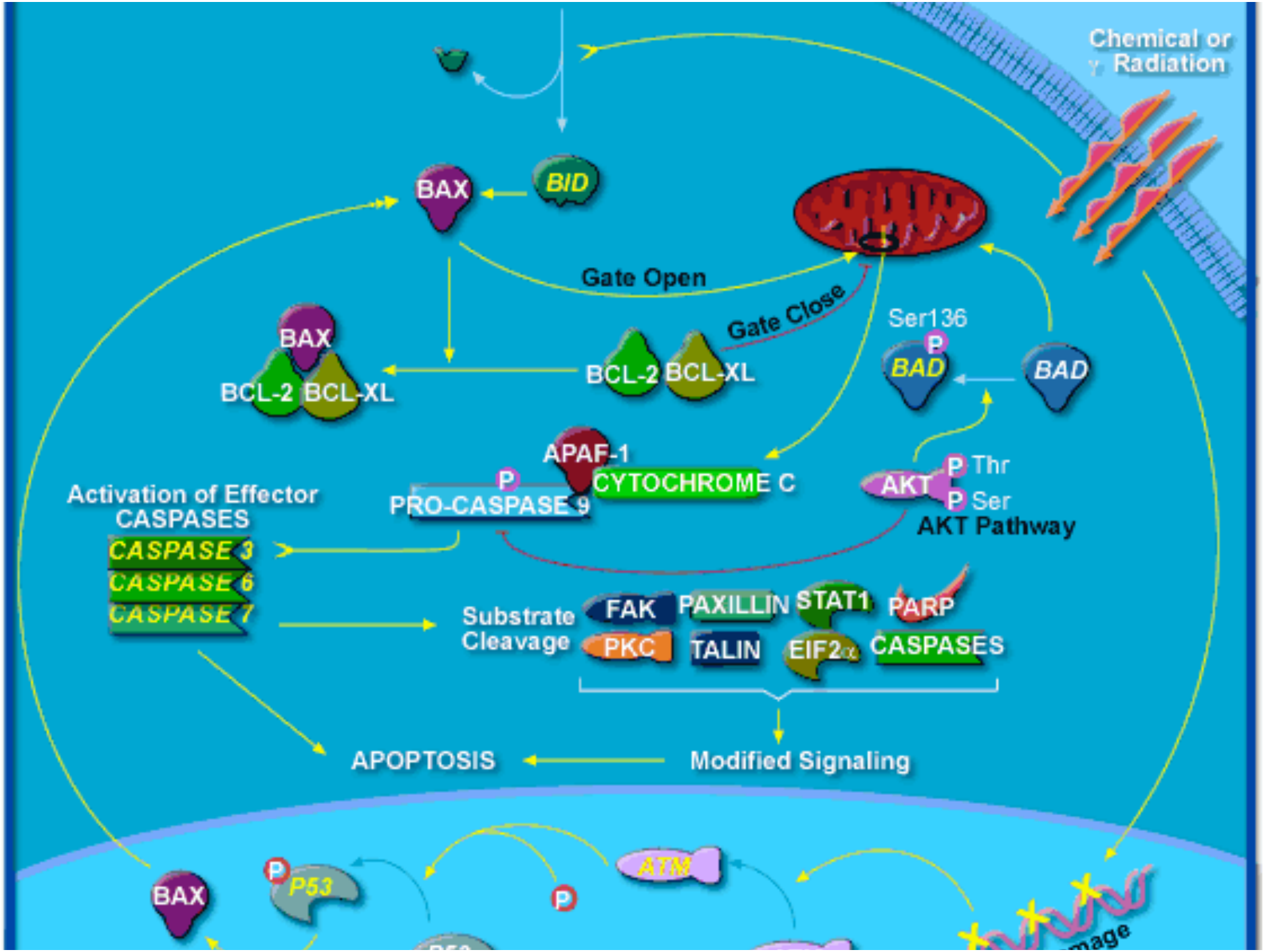
Cancer- 75% of drugs approved- ”standards of care” lack significant impact

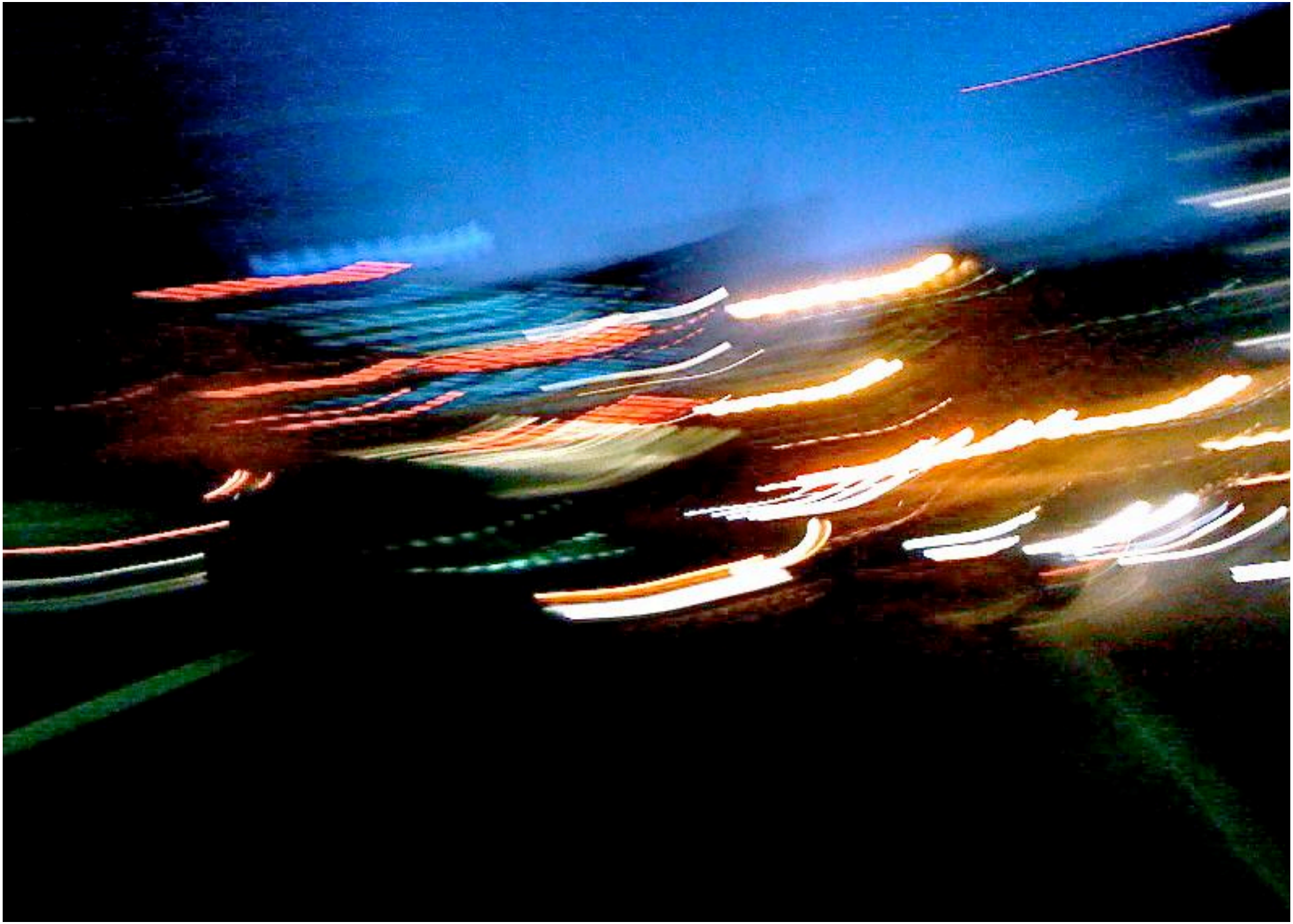
25,000 components with 3,269 associated with disease-yet only hundreds targeted for therapies

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

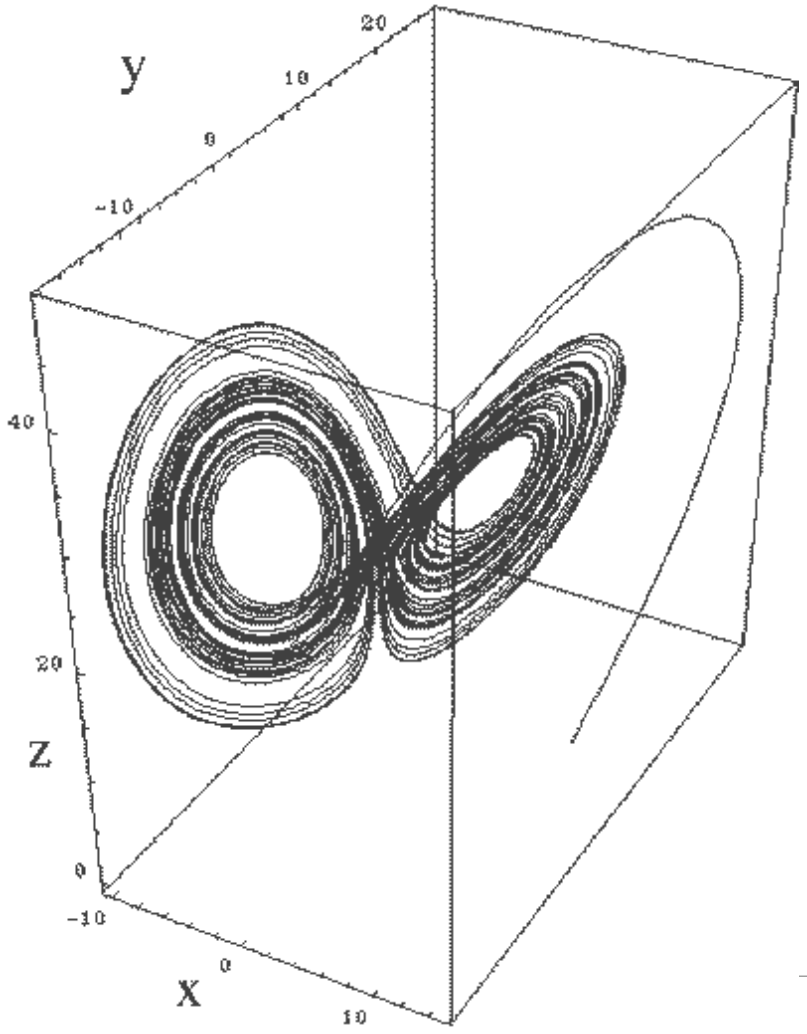
~10% of therapies in Phase I trials will lead to approval

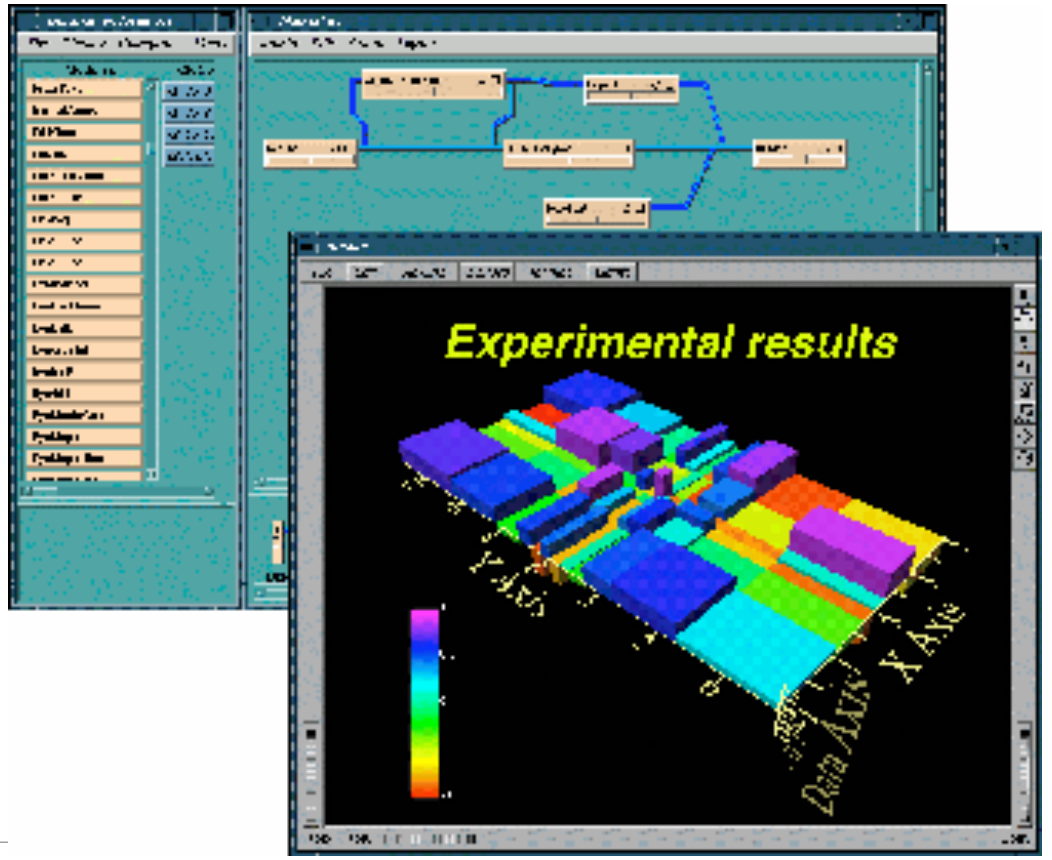
Several specific disease efforts spending ~ $1/3 Billion/year or more to develop therapies

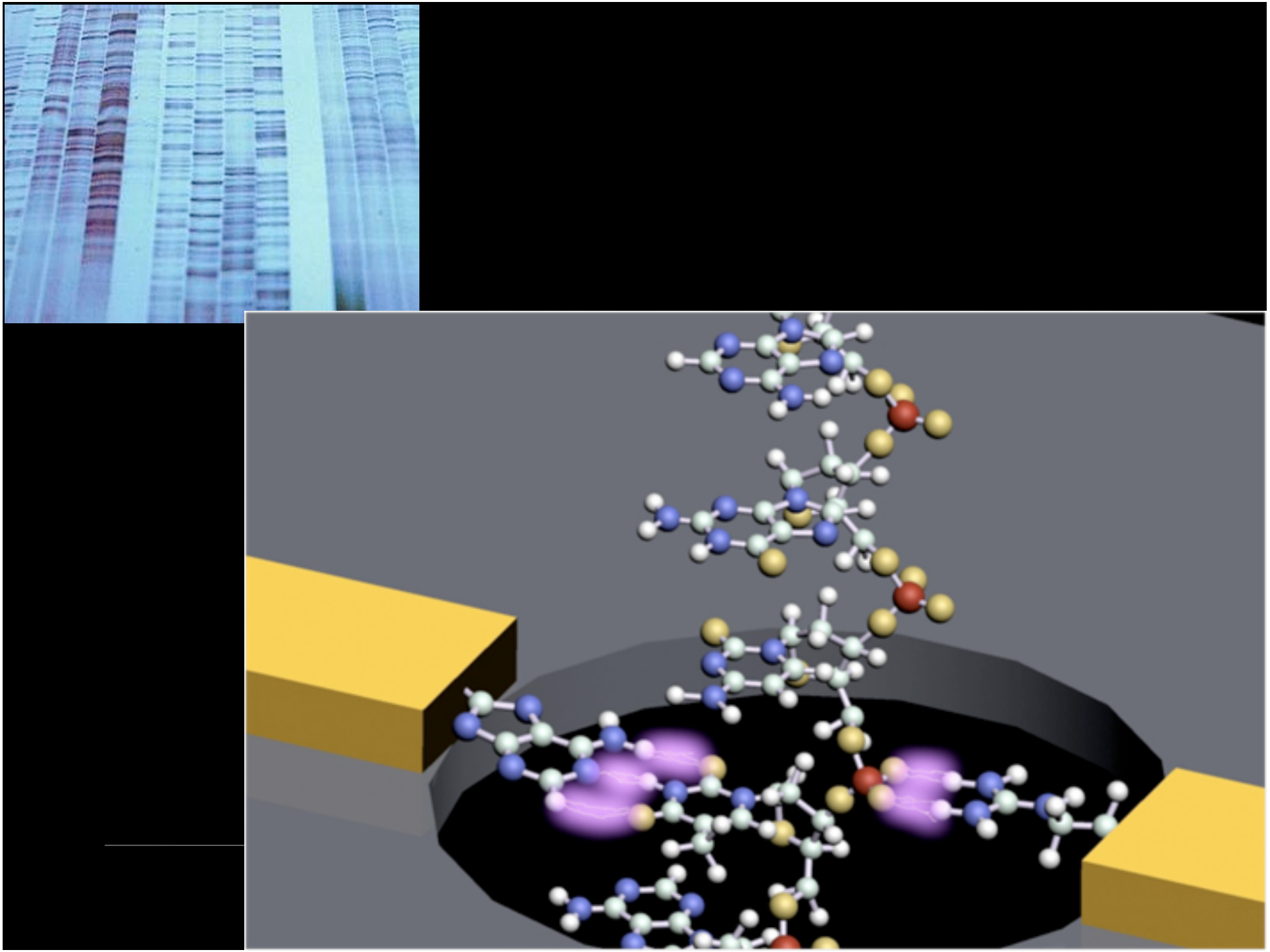

Chemical or
γ Radiation
BAX
BID
Gate Open
Gate Close
BAX
BCL-2 BCL-XL
BCL-2 BCL-XL
Ser136
P
BAD
BAD
APAF-1
CYTOCHROME C
P
PRO-CASPASE 9
AKT
P Thr
P Ser
AKT Pathway
Activation of Effector
CASPASES
CASPASE 3
CASPASE 6
CASPASE 7
Substrate
Cleavage
FAK
PAXILLIN
STAT1
PARP
PKC
TALIN
EIF2α
CASPASES
APOPTOSIS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



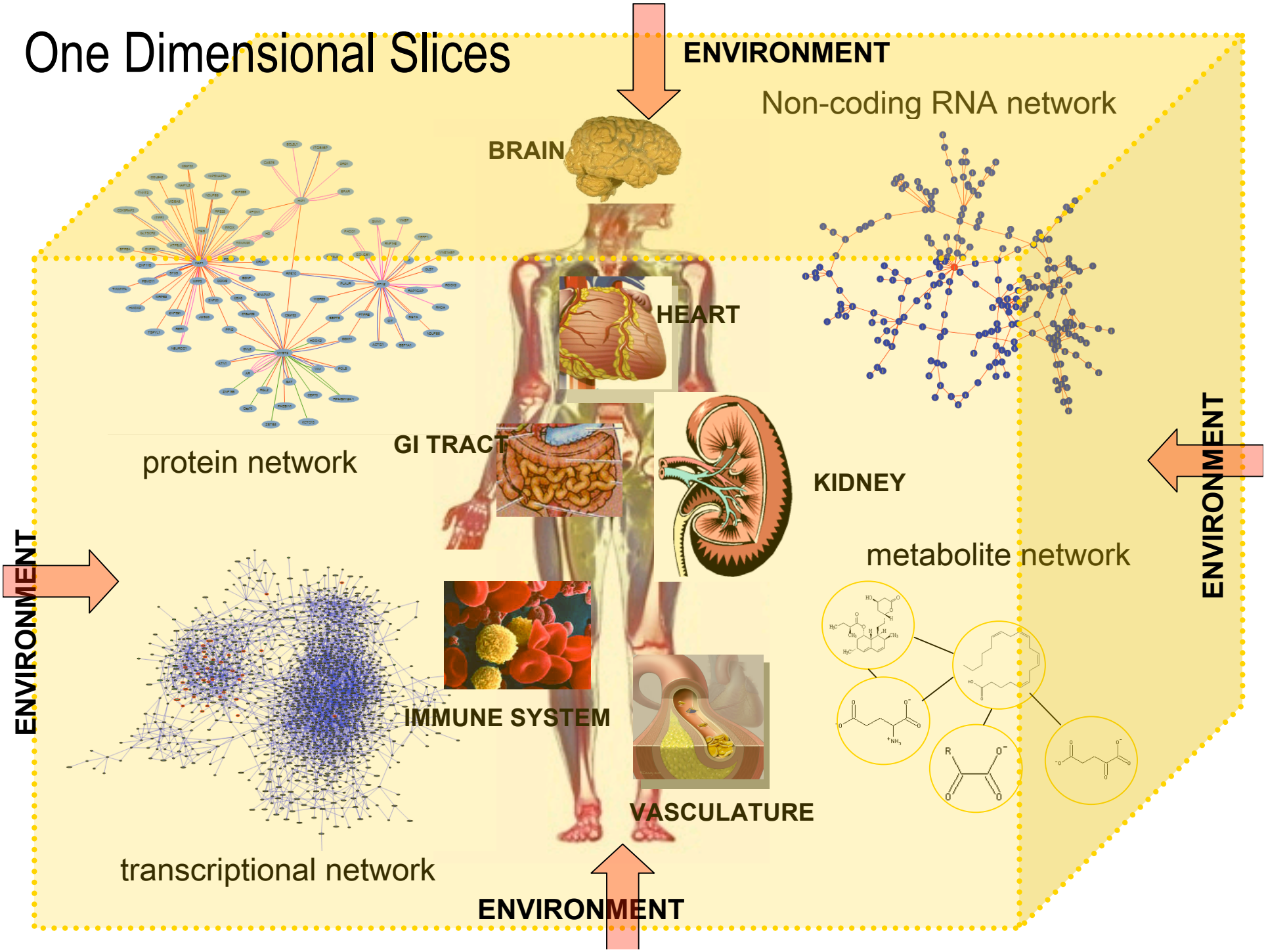
The value of appropriate representations







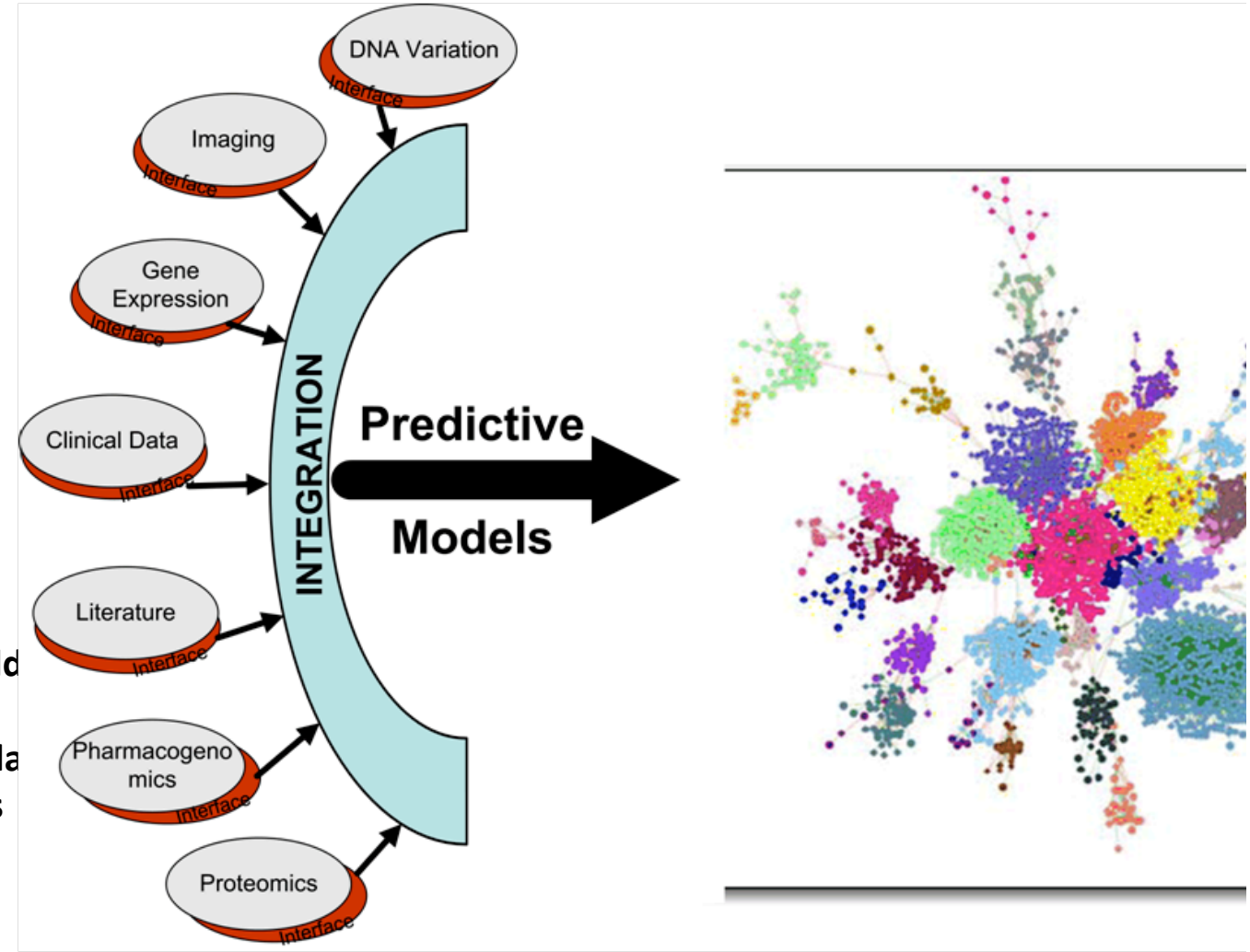
One Dimensional Slices




ORIENS
SEPTEN TRIO
OCCIDENS

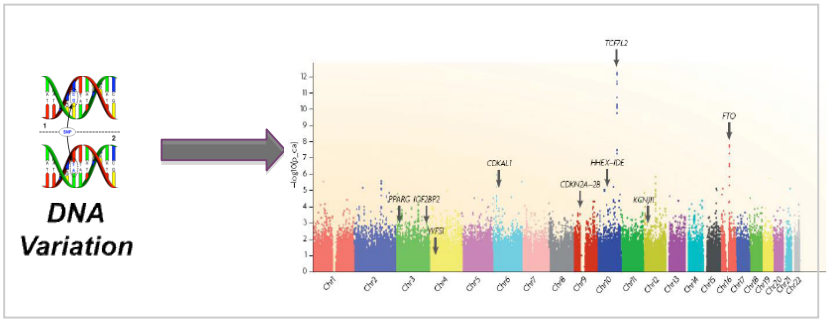
The “Rosetta Integrative Genomics Experiment”: Generation, assembly, and integration of data to build models that predict clinical outcome

5 Year Program
Total Resources
>$150M

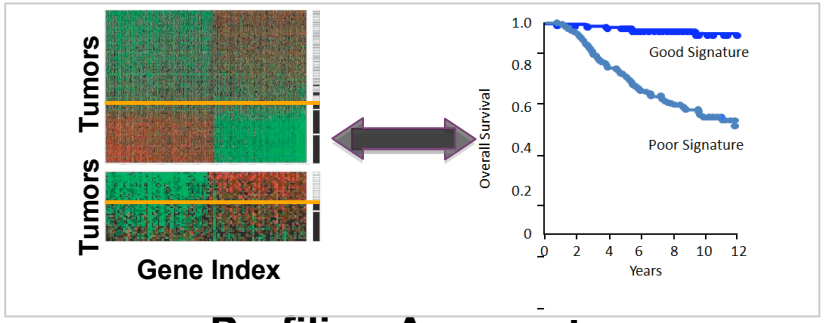


- **Generate data need to build**
- **bionetworks**
- **Assemble other available da**
- **Integrate and build models**
- **Test predictions**
- **Develop treatments**
- **Design Predictive Markers**

How is genomic data used to understand biology?



"Standard" GWAS Approaches

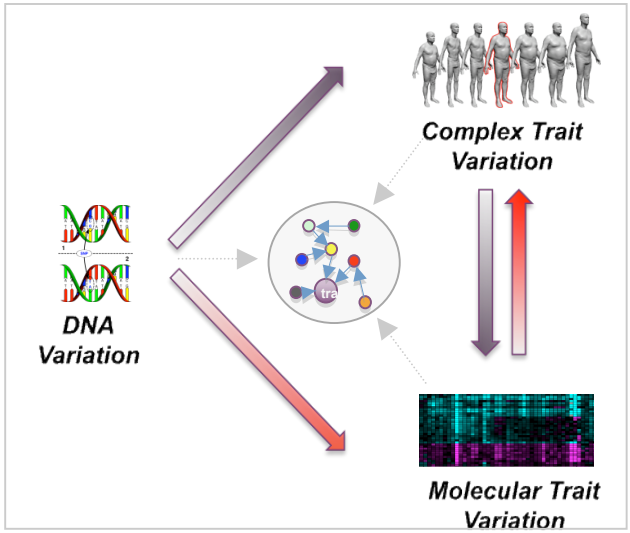


Profiling Approaches

Identifies Causative DNA Variation but provides NO mechanism

Genome scale profiling provide correlates of disease

- **Many examples BUT what is cause and effect?**

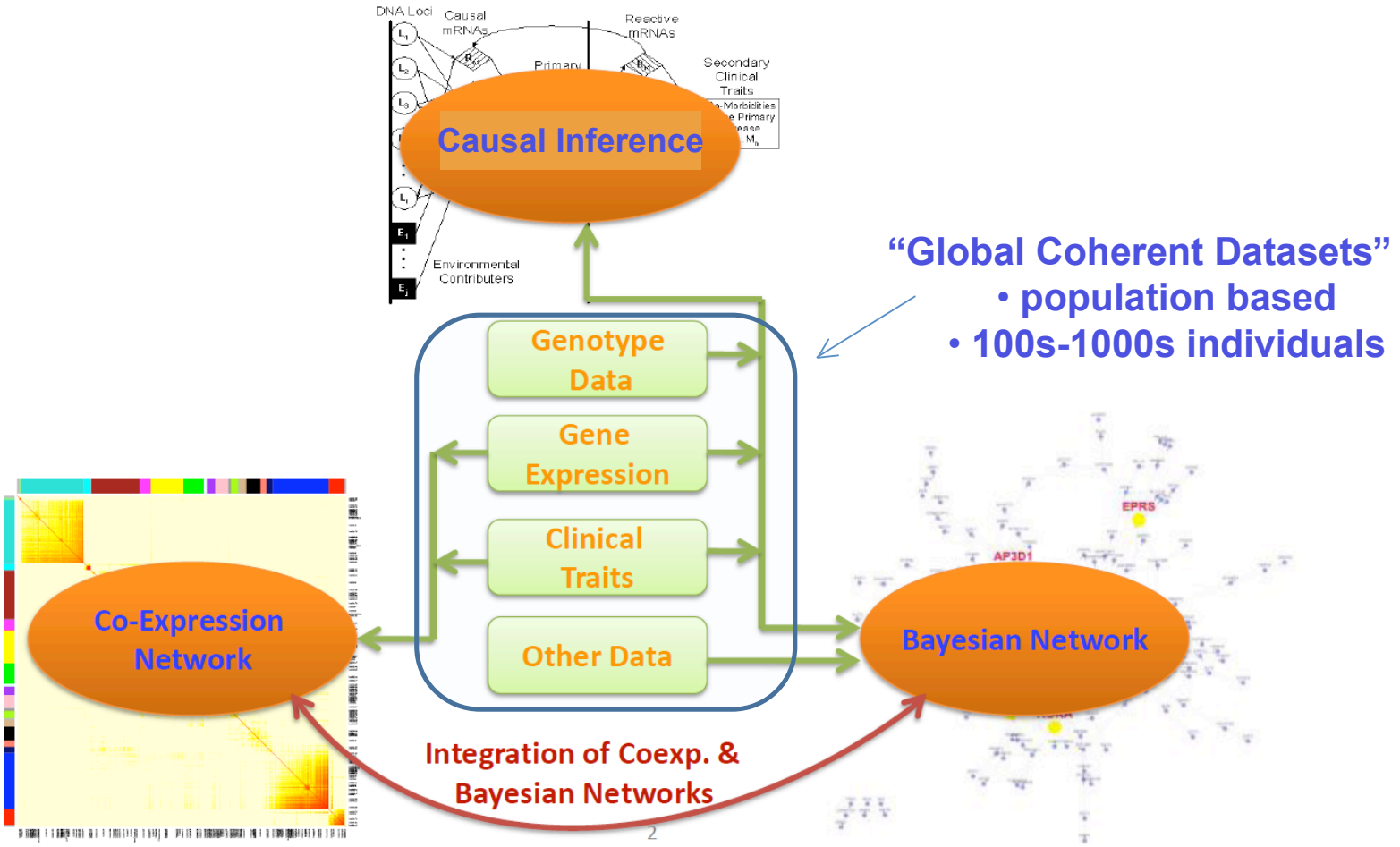


"Integrated" Genetics Approaches

- **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

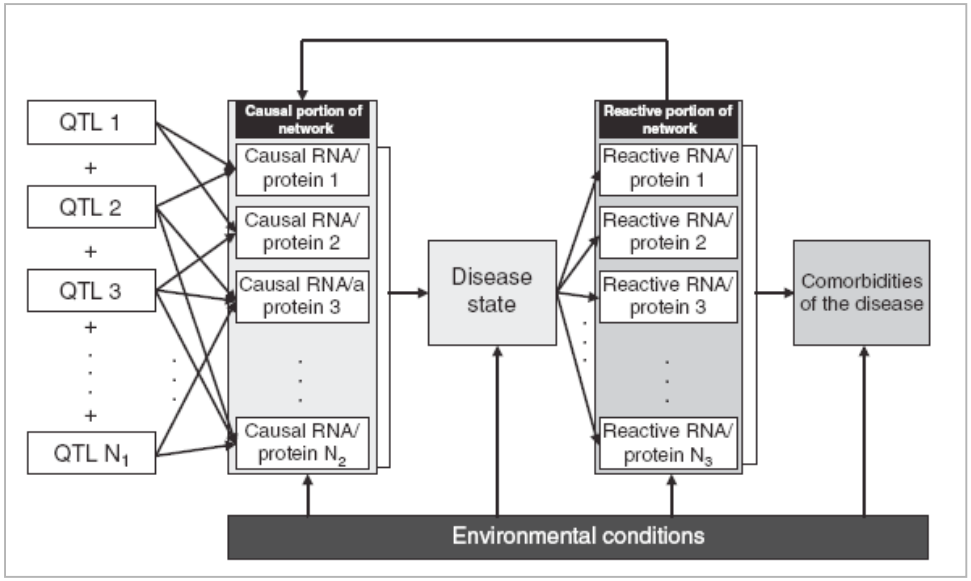
Schadt et al. Nature Genetics 37: 710 (2005)
Millstein et al. BMC Genetics 10: 23 (2009)



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

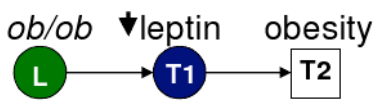
Zhu et al. Cytogenet Genome Res. 105:363 (2004)
Zhu et al. PLoS Comput. Biol. 3: e69 (2007)

Causal Inference



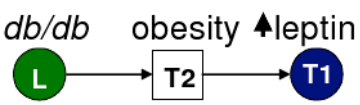
Start with this information

Causative Model



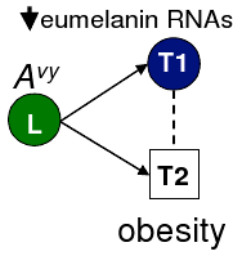
$$P(L,T_1,T_2) = P(T_1 \mid L)P(T_2 \mid T_1)$$

Reactive Model



$$P(L,T_1,T_2) = P(T_2 \mid L)P(T_1 \mid T_2)$$

Independent Model



$$P(L,T_1,T_2) = P(T_2 \mid L)P(T_1 \mid L)$$

L DNA Locus controlling RNA levels and/or clinical traits

T1 Quantitative trait 1

T2 Quantitative trait 2

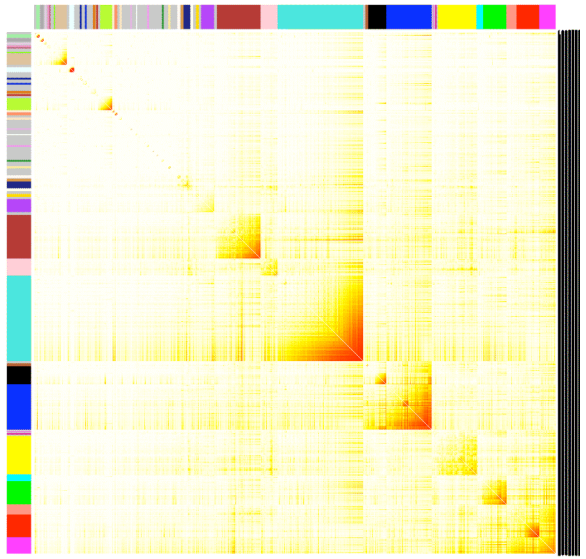
Nature Genetics: **37(7)**:710-717

Co-expression Networks

Co-expression/Association Networks

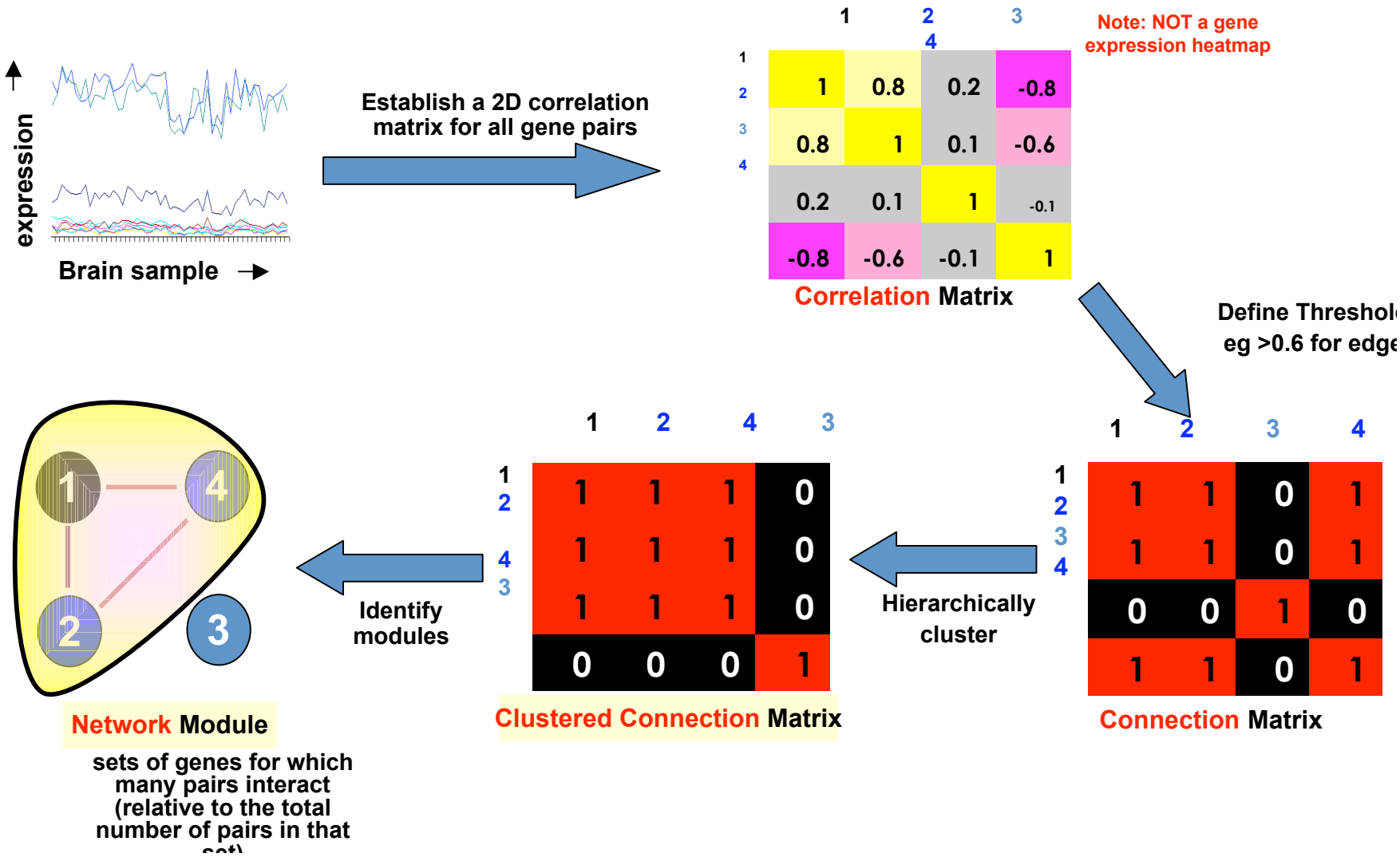
(does Gene A associate with Gene B?)

- **Correlation-based associations**
- **Protein-protein interactions**
- **Literature-based associations**
- **Knowledge-based (KEGG, GO, etc.)**
- **DESCRIPTIVE**: ***CAN NOT be used to predict outcomes or perturbations***



Constructing Co-expression Networks

Start with expression measures for ~13K genes most variant genes across 100-150 samples

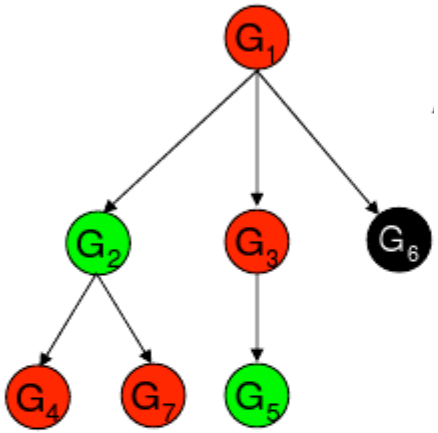


Bayesian Networks

Bayesian Networks

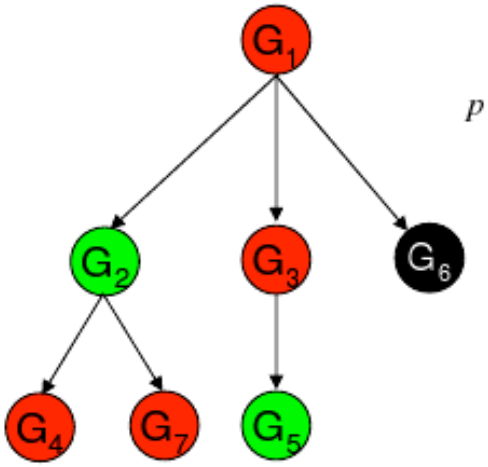
(does Gene A control Gene B?)

- **Captures the stochastic nature of biological system**
- **Probability based**
- **Can include priors such as causality information**
- **PREDICTIVE: *CAN be used to predict outcomes or perturbations***



Sage BIONETWORKS

Constructing Bayesian Networks



Down regulated

Up regulated

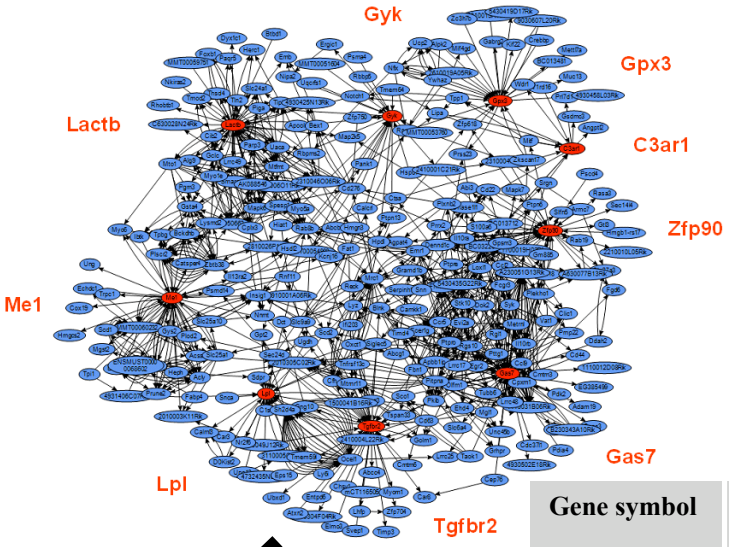
Not changed

- BN method provides a way to decompose a joint probability distribution based on conditional independence

$$p(Network) = p(G_1, G_2, G_3, G_4, G_5, G_6, G_7)$$
$$= p(G_1)p(G_2 \mid G_1)p(G_3 \mid G_1)p(G_6 \mid G_1)p(G_4 \mid G_2)p(G_7 \mid G_2)p(G_5 \mid G_3)$$

- For a given network, we find the maximum likelihood of the network given the observed data *D, p(D|Network)*
- Training Bayesian Networks
  - We want to search the space of all networks to find the optimal one
  - Calculate probability tables associated with the networks
- To find the best network we perform the search 1,000 times using random seeds
  - Computationally intense procedure
  - Presently runs on a 6000+ CPU (IBM Blade) Cluster
- Common features are then extracted (e.g., connections seen in > 30% of the networks are extracted) and probability tables are updated

Preliminary Probabalistic Models- Rosetta /Schadt



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

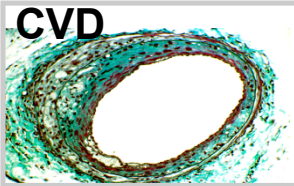
Gene symbol	Gene name	Variance of OFPM explained by gene expression*	Mouse model	Source
Zfp90	Zinc finger protein 90	68%	tg	Constructed using BAC transgenics
Gas7	Growth arrest specific 7	68%	tg	Constructed using BAC transgenics
Gpx3	Glutathione peroxidase 3	61%	tg	Provided by Prof. Oleg Mirochnitchenko (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**

"*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.**

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

- **#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.

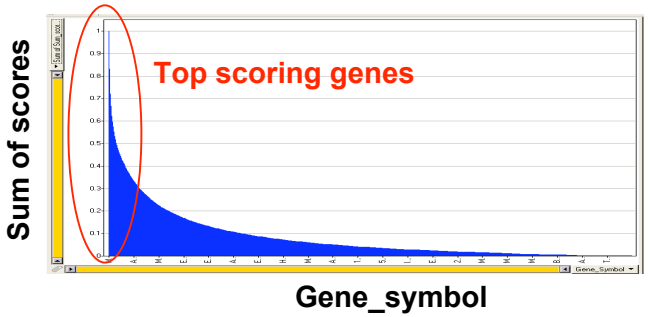
- **#5 - Define clinically relevant subpopulations**
  - Workflow - Similar to #4 above, but typically starting at an earlier stage to incorporate hypotheses about population segments early enough in the development process that they are easily tested prospectively.

- **#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.

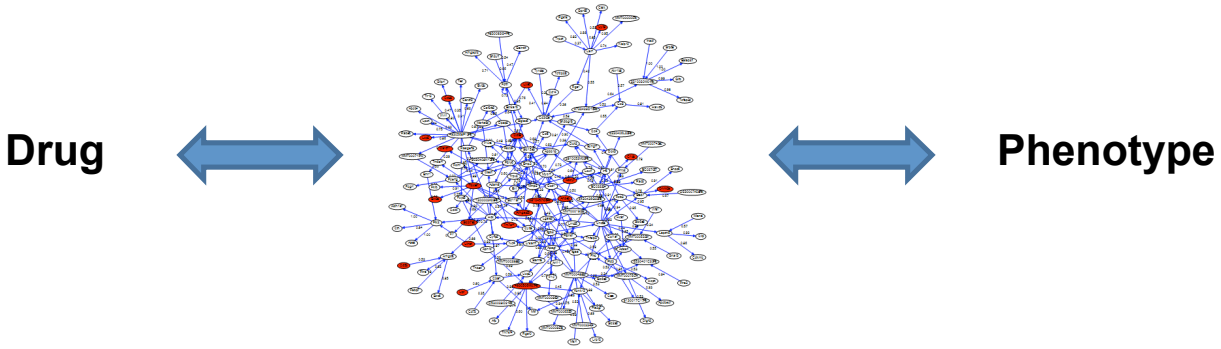
#3 Repositioning Strategy

Two approaches to match a “safe” compound to a phenotype

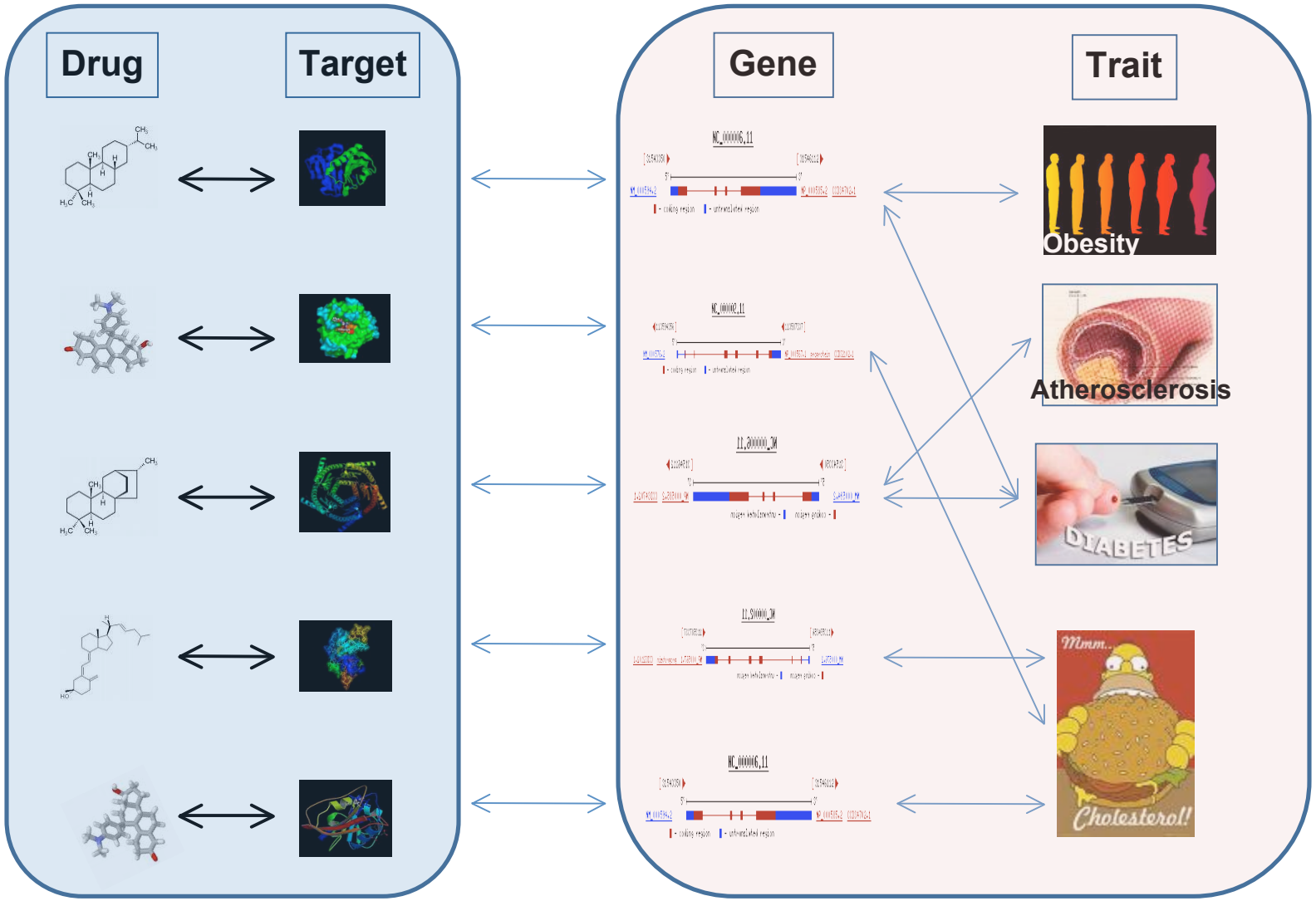
Approach 1: Search network maps that causally link targets to phenotypes for both mouse and human



Approach 2: Map compound signatures to networks that are linked to phenotypes for both mouse and human



Approach 3.1: Link Drug/Target data to Target/Trait data

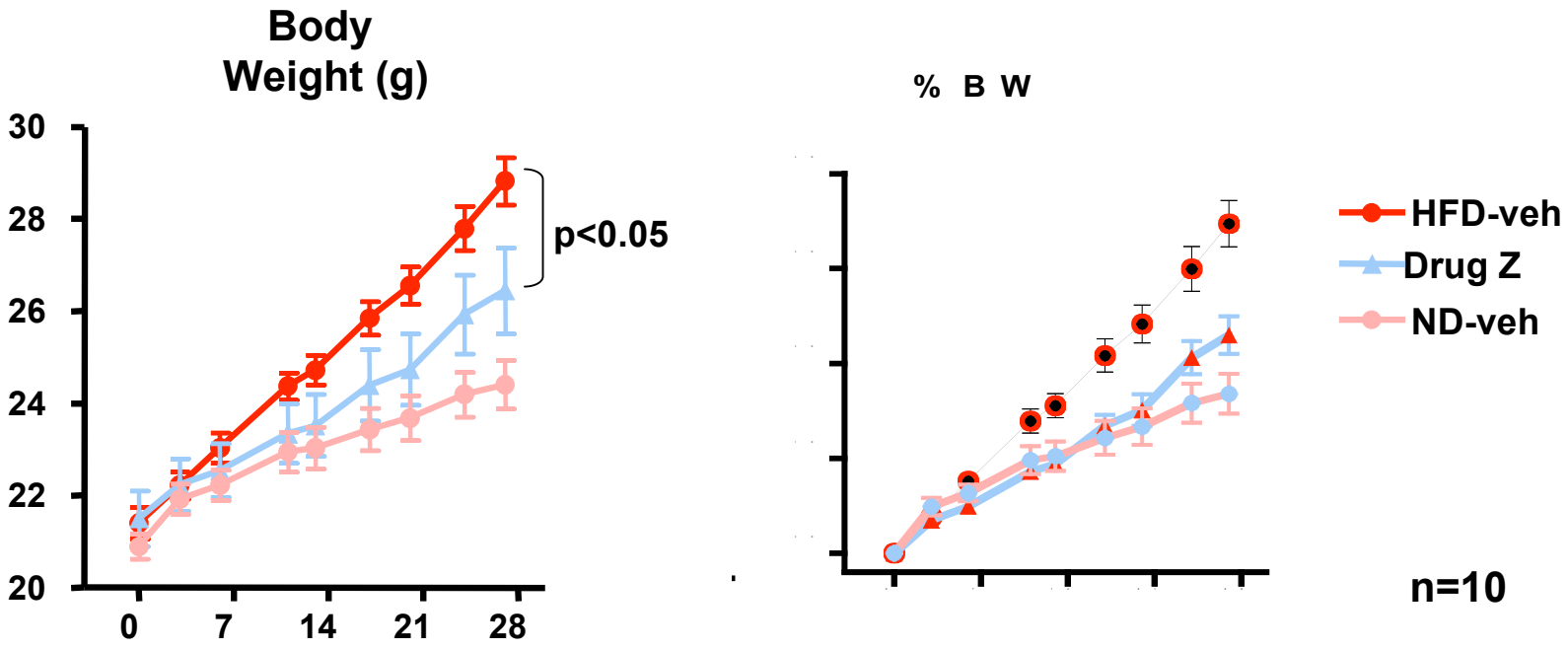


Public and/or Pharma Proprietary Data
Links drugs to targets

Sage Data
Links genes to traits

Preclinical Pharmacologic Validation of Drug Z

Drug developed for another indication showing evidence for association to obesity traits in mouse F2 crosses and being validated pharmacologically in preclinical model of HFD feeding



- **~35% reduction in body weight gain (*vs* vehicle).**
- **also reductions in Leptin & insulin in DIO model.**

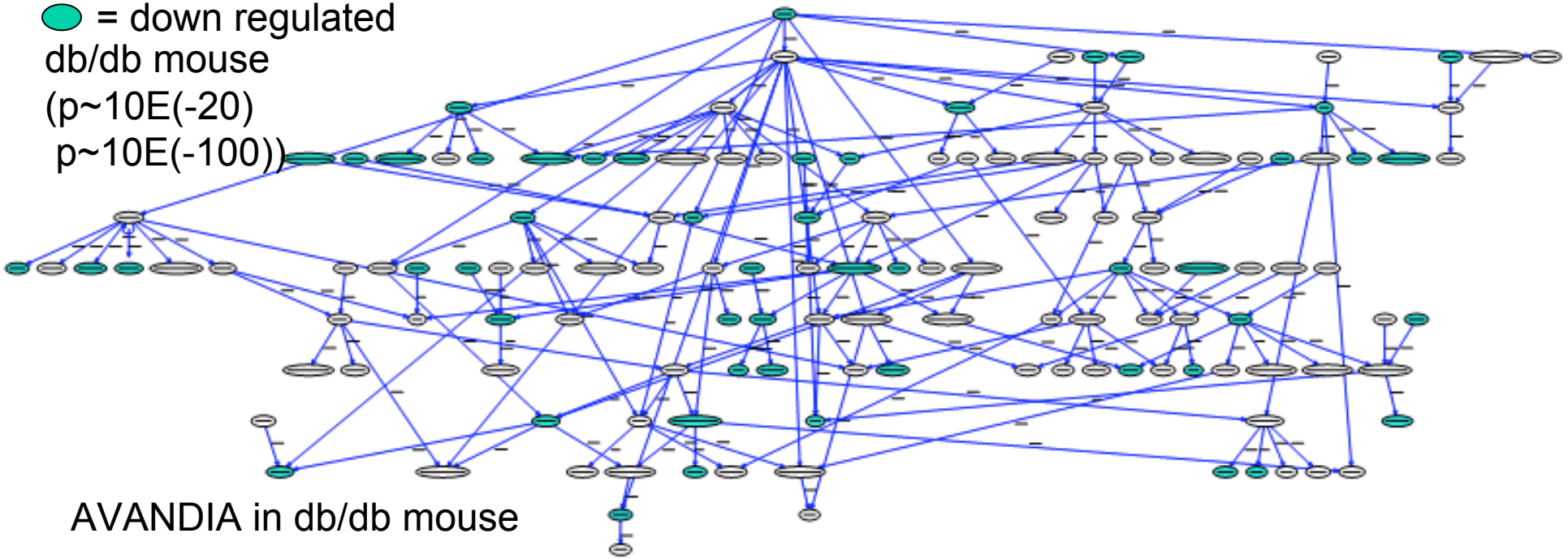
Ability to integrate compound data into our network analyses

db/db mouse
(p~10E(-30))

= up regulated

= down regulated

db/db mouse
(p~10E(-20)
p~10E(-100))



AVANDIA in db/db mouse

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

Ideker: Assembling Networks for Use in Clinic

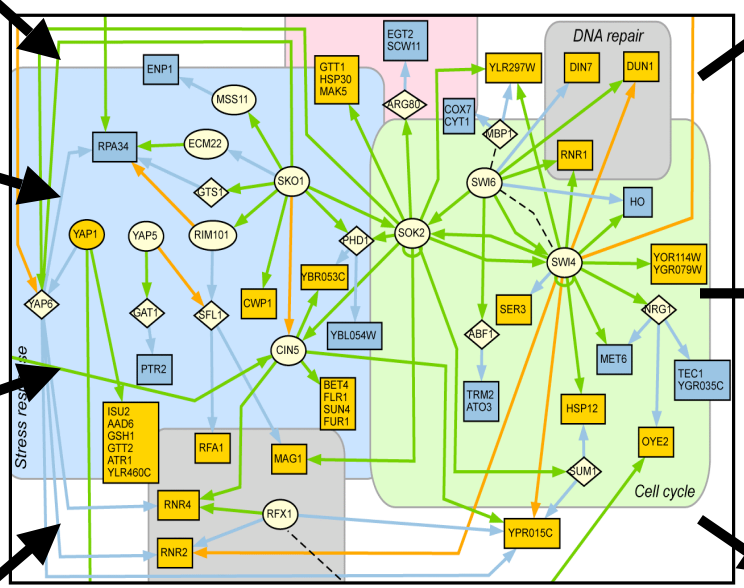
Network evolutionary comparison / cross-species alignment to identify conserved modules

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

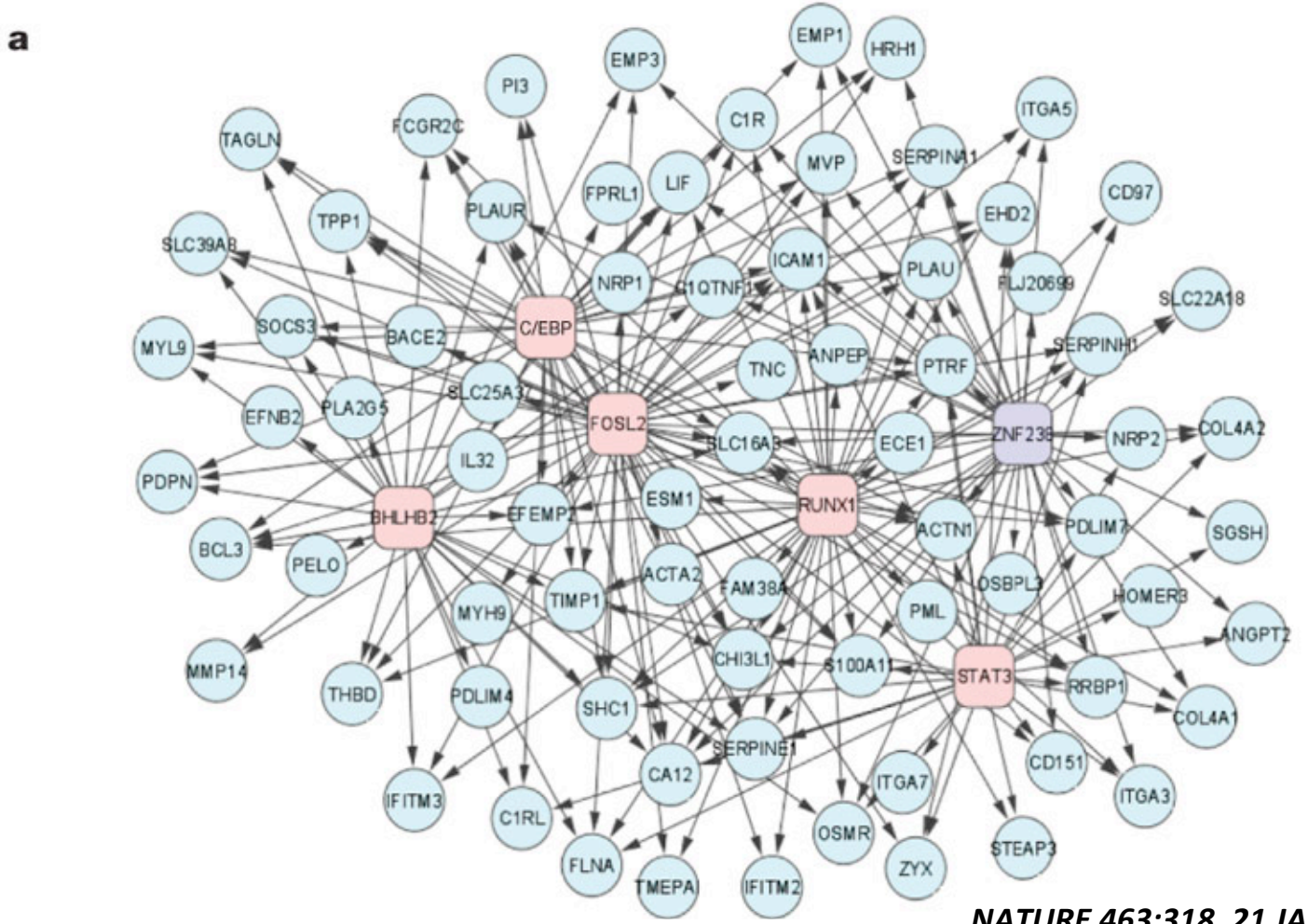
Identification of networks associated with cancer progression

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

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

The transcriptional network for mesenchymal transformation of brain tumours

Maria Stella Carro[1]*{, Wei Keat Lim[2,3]*{, Mariano Javier Alvarez[3,4]*, Robert J. Bollo[8], Xudong Zhao[1], Evan Y. Snyder[9], Erik P. Sulman[10], Sandrine L. Anne[1]{, Fiona Doetsch[5], Howard Colman[11], Anna Lasorella[1,5,6], Ken Aldape[12], Andrea Califano[1,2,3,4] & Antonio Iavarone[1,5,7]



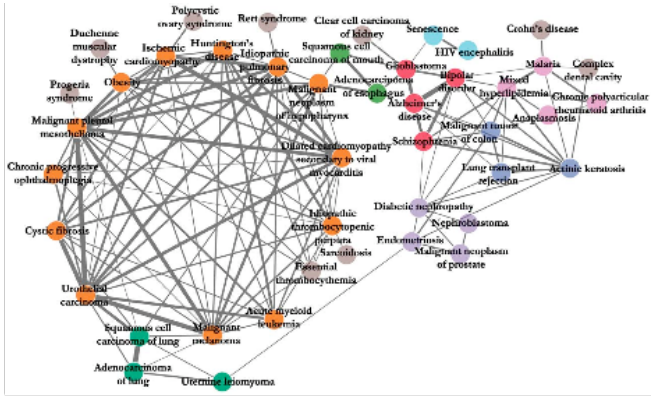
NATURE 463:318, 21 JANUARY 2010

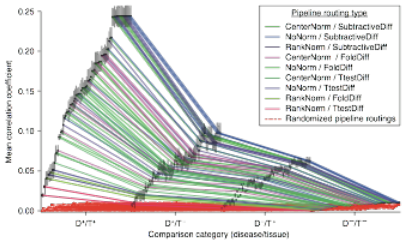
Atul Butte Stanford

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



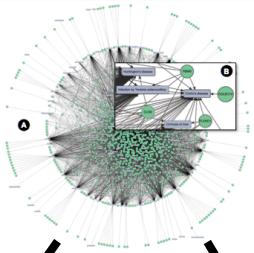


Joel Dudley et al.. Molecular systems biology (2009) vol. 5 pp. 307

High quality signals exist in public data

Differences ⟷ Commonalities

Plasma proteome networks

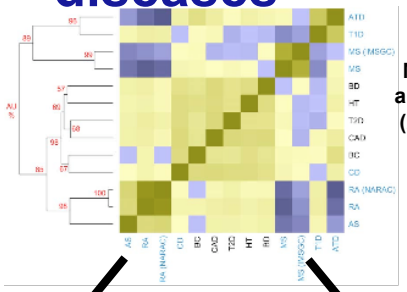


Joel Dudley and Atul Butte. Pacific Symposium on Biocomputing (2009) pp. 27-38

Which biomarkers best discriminate diseases?

Is there a blood biomarker for general pathology?

Genetic architecture of autoimmune diseases

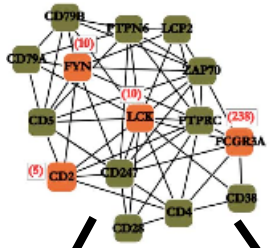


Marina Sirota et al. PLoS genetics (2009) vol. 5 (12) pp. e1000792

Are there genetic "switches" for autoimmunity?

Is there a common autoimmune susceptibility variant?

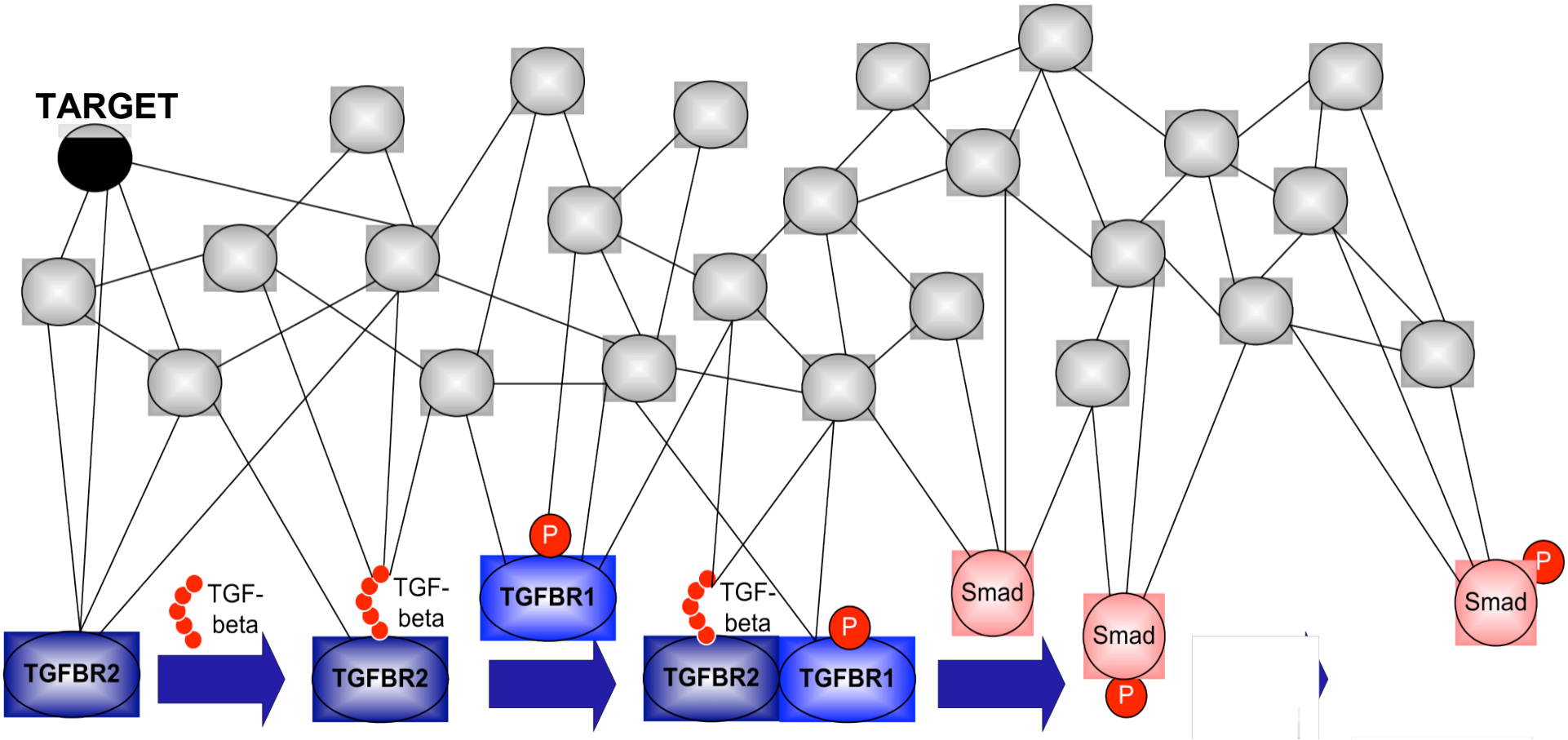
Functional gene module networks



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?


TARGET
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, 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

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

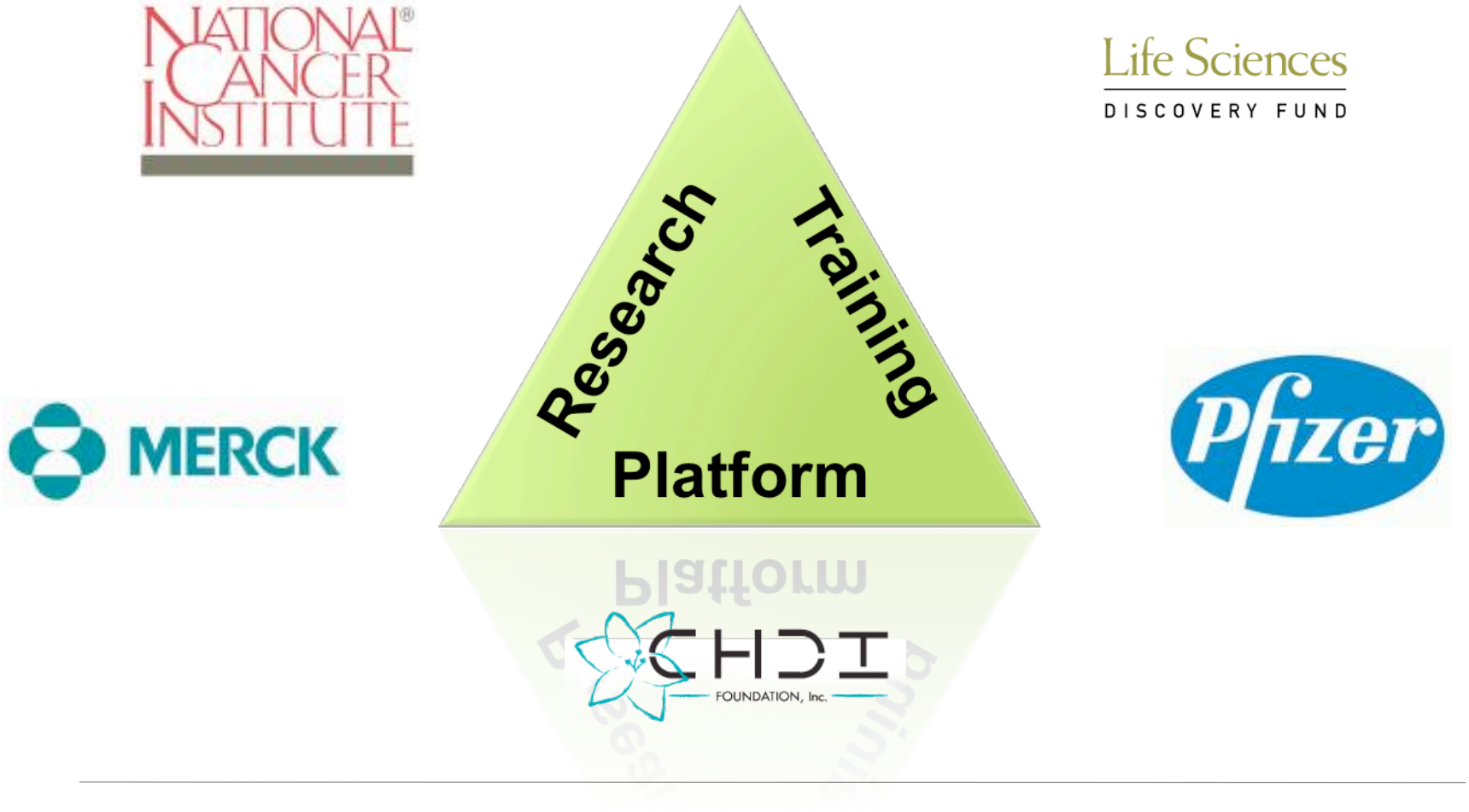
Sage Bionetworks is a non-profit organization with a vision to create a “commons” where integrative bionetworks are evolved by contributor scientists with a shared vision to accelerate the elimination of human disease

Sage Bionetworks

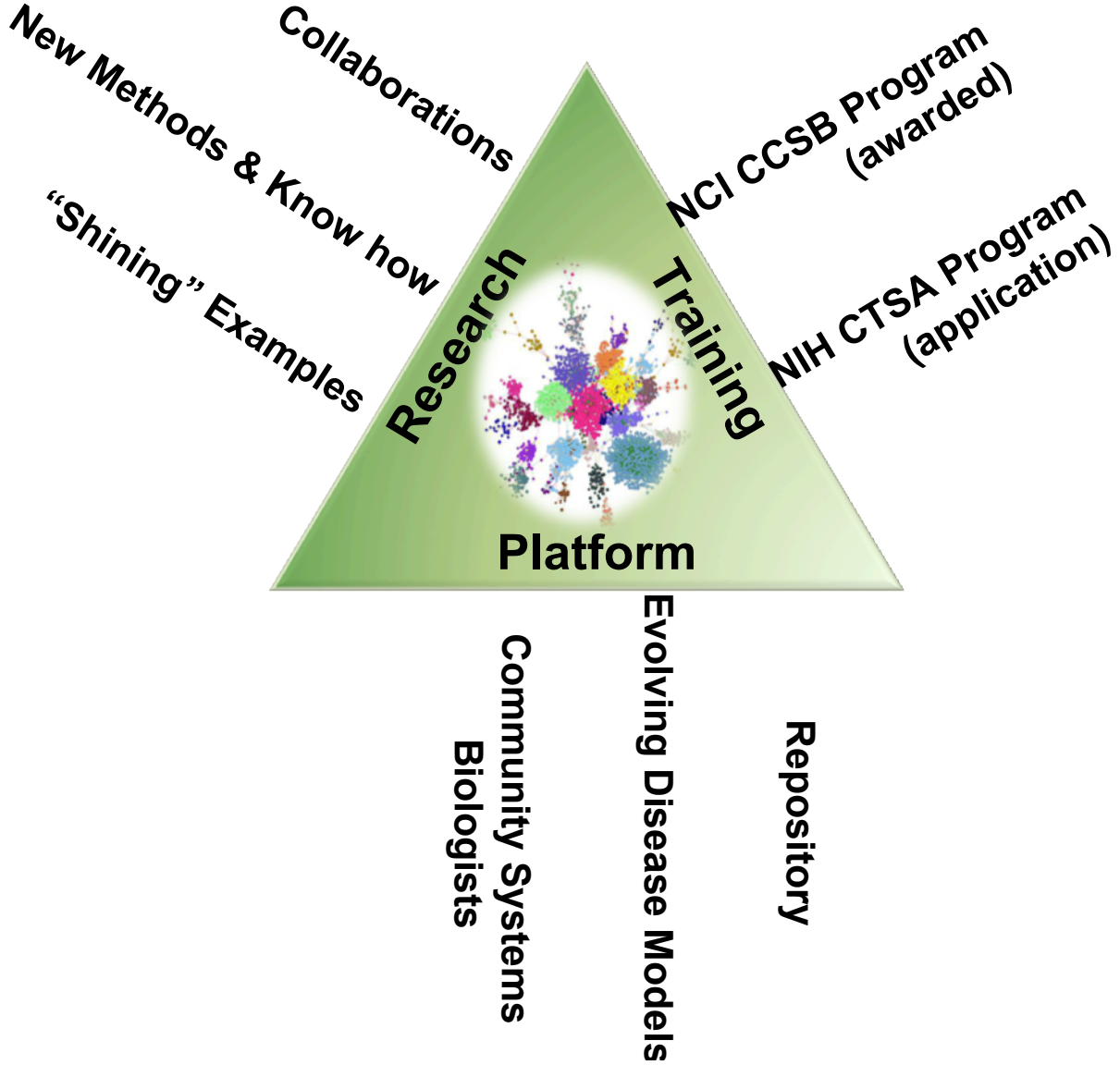
FRED HUTCHINSON CANCER RESEARCH CENTER

A LIFE OF SCIENCE

Sage Bionetworks

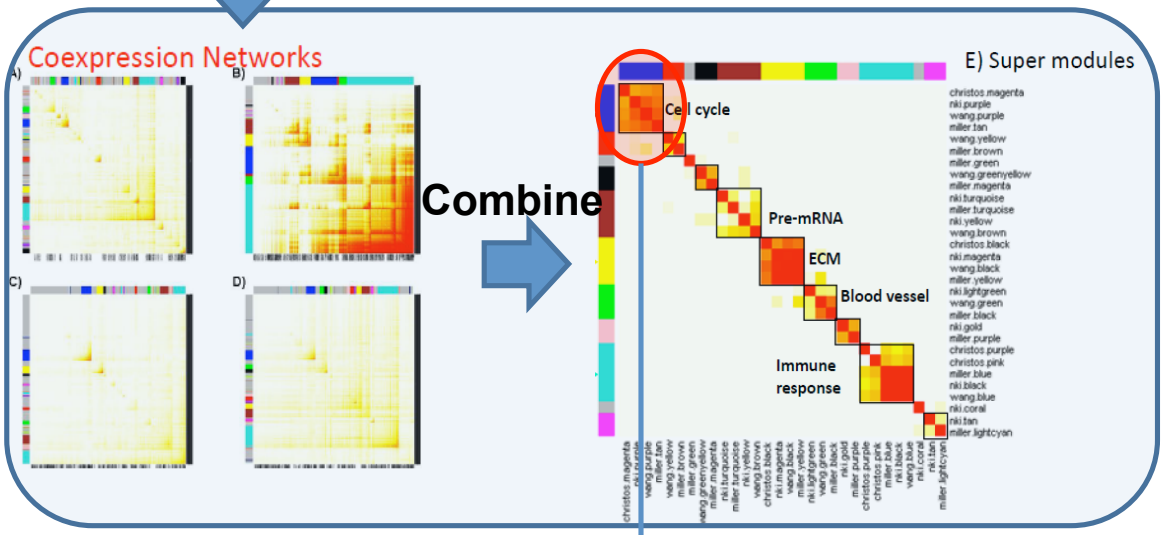


Sage Bionetworks

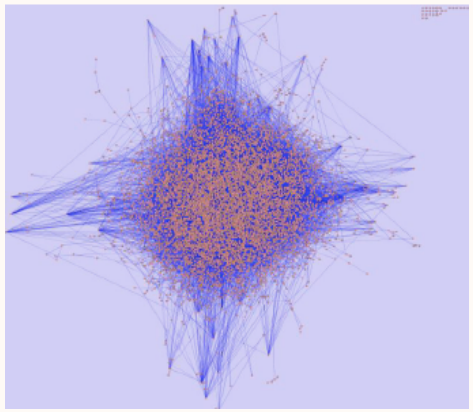


Example 1: Identification of Molecular Drivers of Breast Cancer

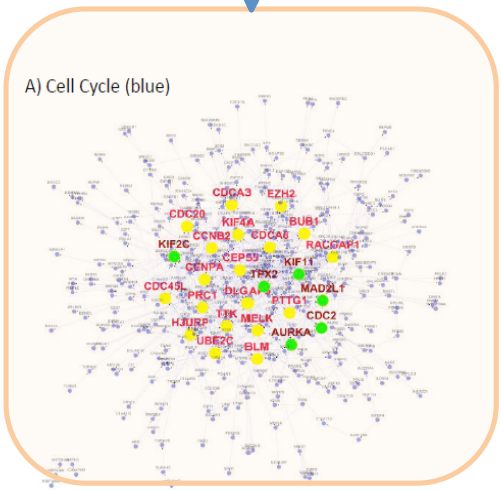
4 Published Breast Cancer Gene Expression Studies



Co-expression Network

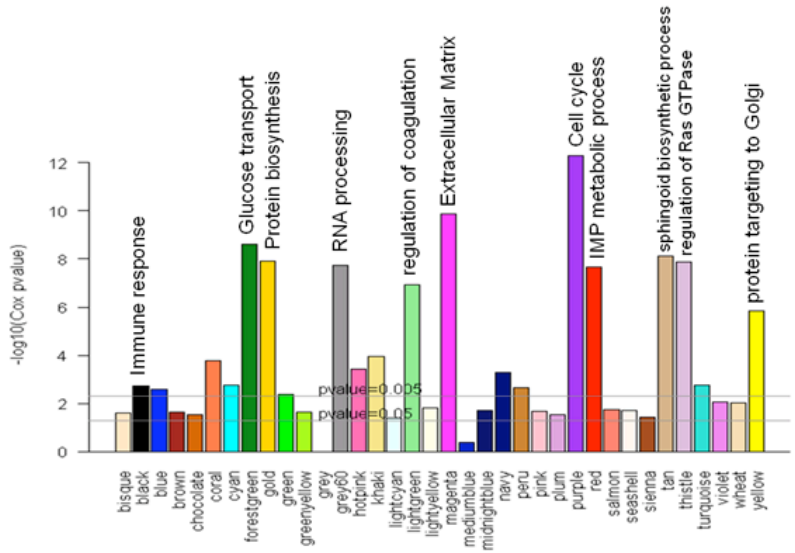


Bayesian Network

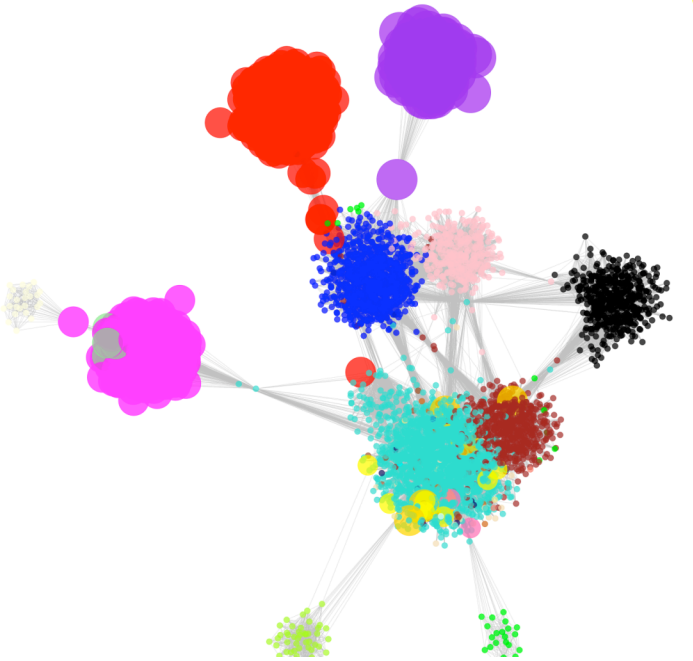


Bayesian Sub-network with Global Drivers (yellow)

Co-expression sub-networks predict survival

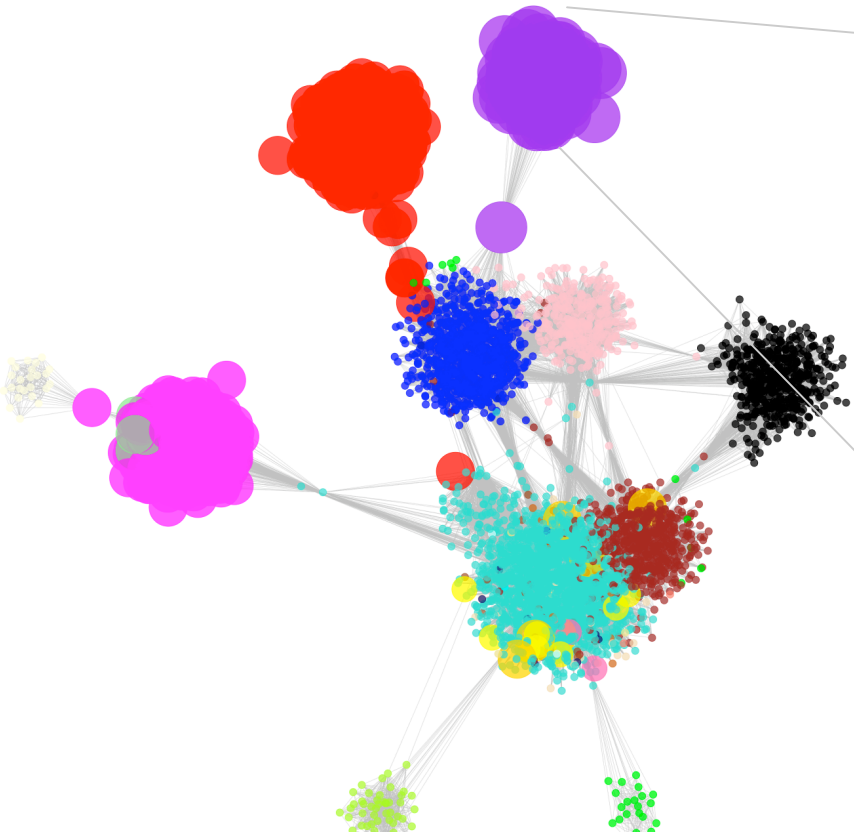


Prognostic power of the gene modules in the NKI gene co-expression network. Module prognostic power was defined as -log(Cox p-value) from a multi-variate Cox proportional hazards regression model that regresses patient survival onto the principal components of a given module

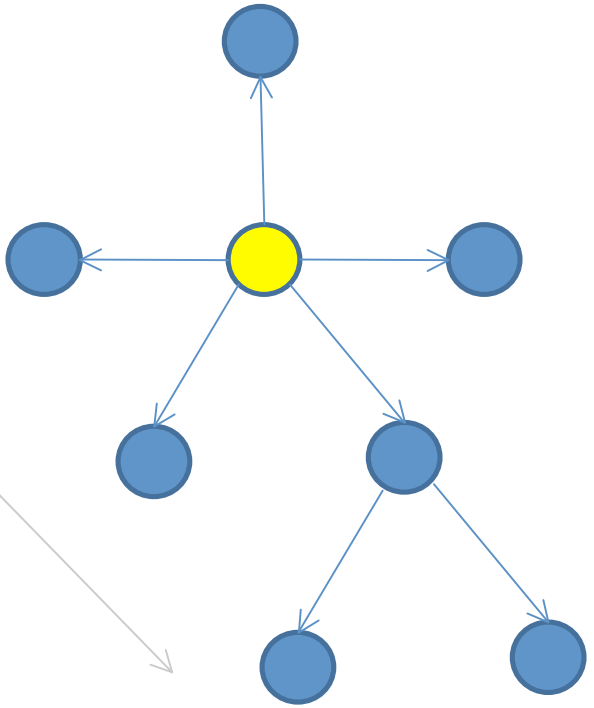


NKI Co-expression network

Network Properties can be used to predict key drivers



NKI Co-expression network

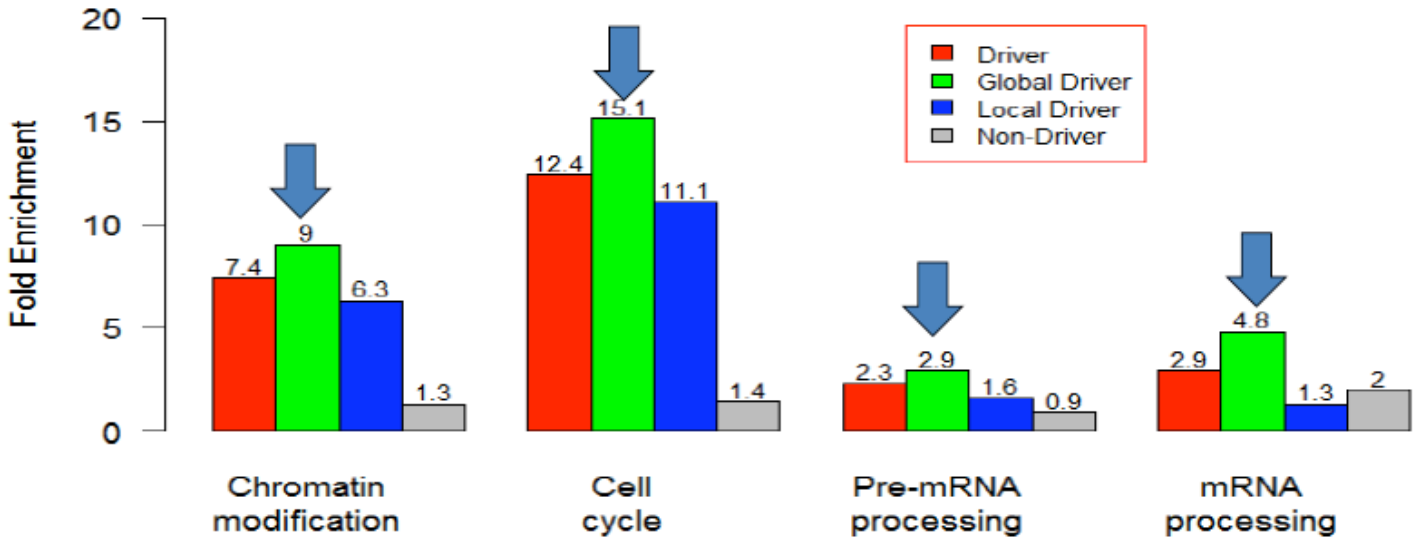


For a given BN, let μ be the numbers of N-hub downstream nodes and d be the outdegrees for all the genes. Genes with the number of N-hub downstream nodes greater than mean (μ) + sd (μ) are nominated as regulators.

Drivers from Network are More likely to have Survival Effect in siRNA Screens

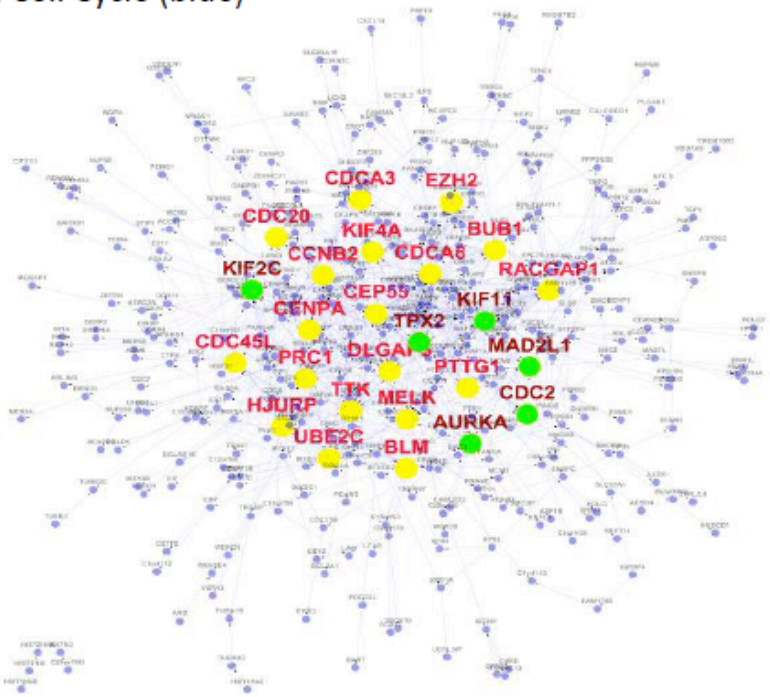
- Cell Cycle siRNA Signature (CCSS)

- 210 genes from an siRNA library targeting 2,500 cell cycle genes were identified to be important to cell survival

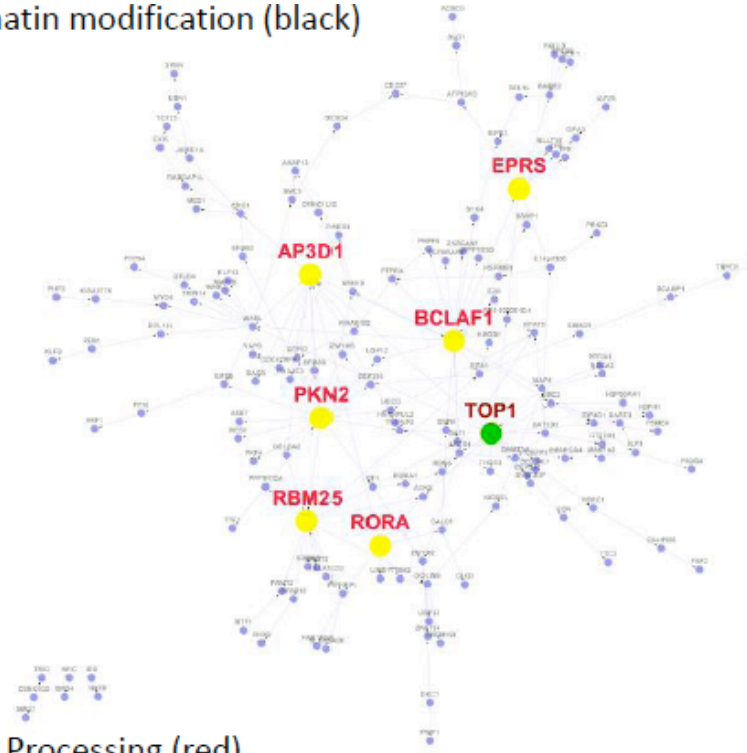


Key Drivers

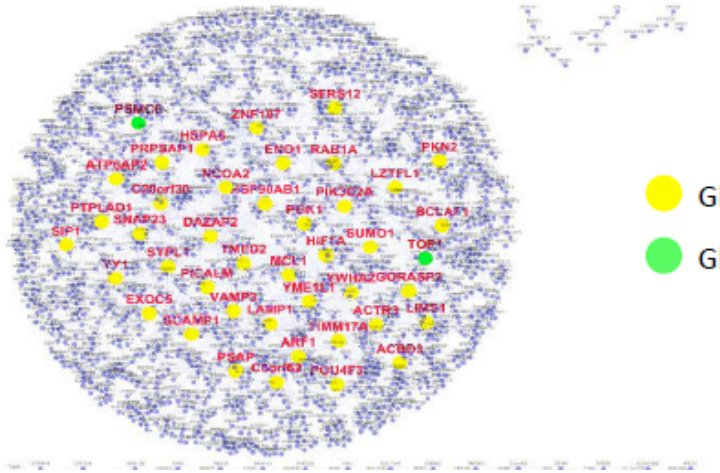
A) Cell Cycle (blue)



B) Chromatin modification (black)



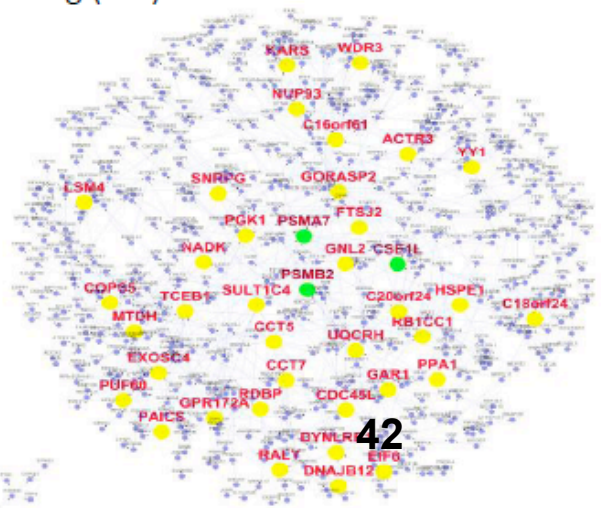
C) Pre-mRNA Processing (brown)



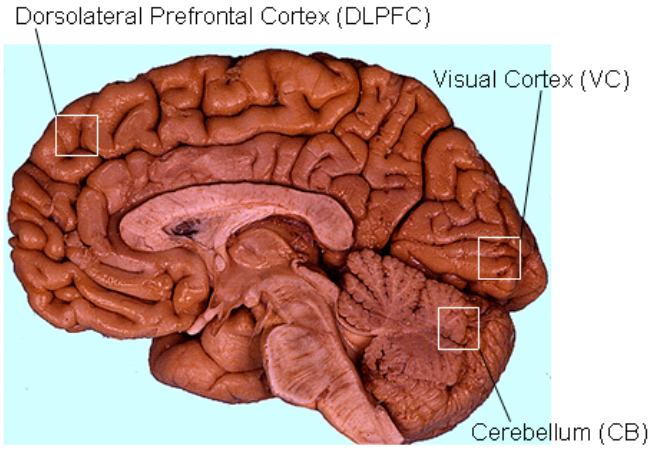
Global driver

Global driver & CCSS

D) mRNA Processing (red)



Application II: Alzheimer's Disease

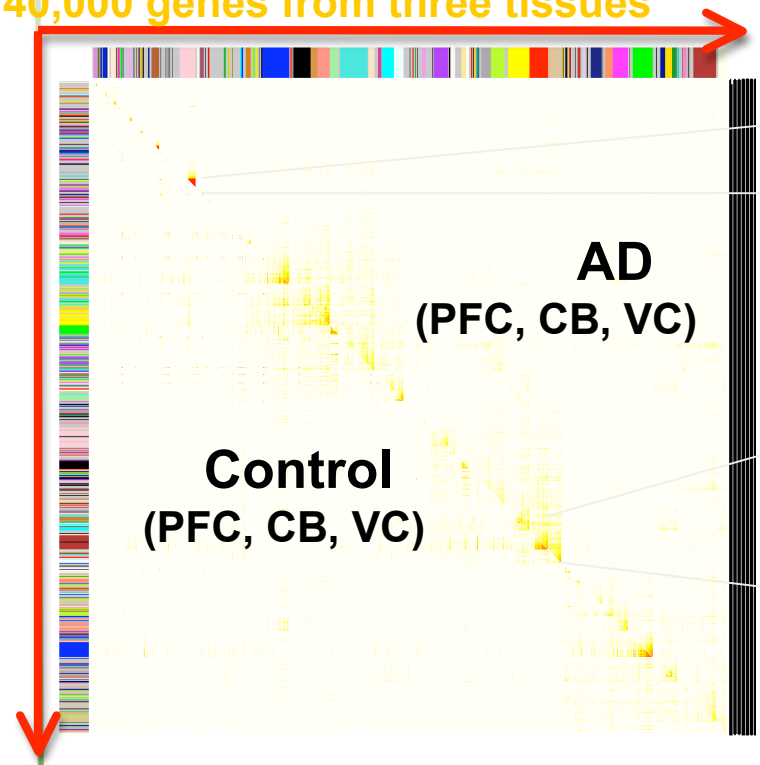


- Cross-tissue coexpression networks for both normal and AD brains
  - prefrontal cortex, cerebellum, visual cortex
- Differential network analysis on AD and normal networks
- Integrate coexpression networks and Bayesian networks to identify key regulators for the modules associated with AD

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

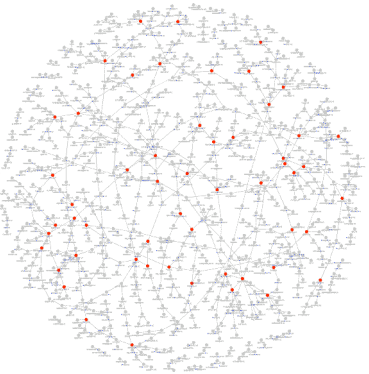
Identification of Disease (AD) Pathways via Comparative Gene Network Analysis

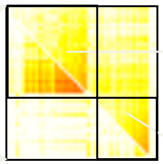
40,000 genes from three tissues



Glutathione transfer

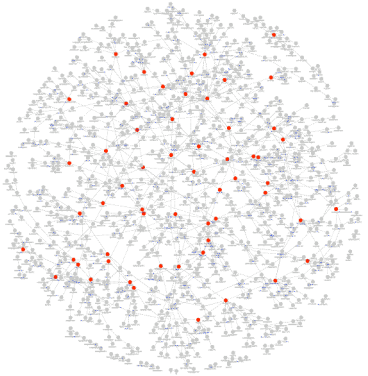
Gain connectivity by 91 fold





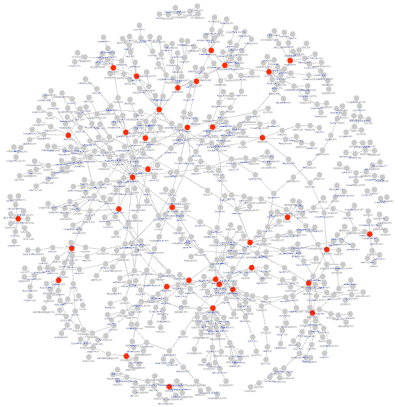
nerve ensheathment

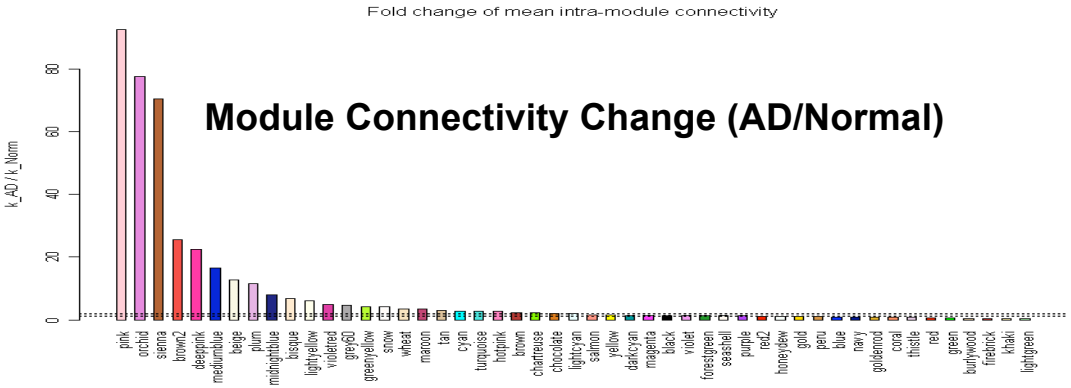
Lose connectivity by 40%



extracellular matrix

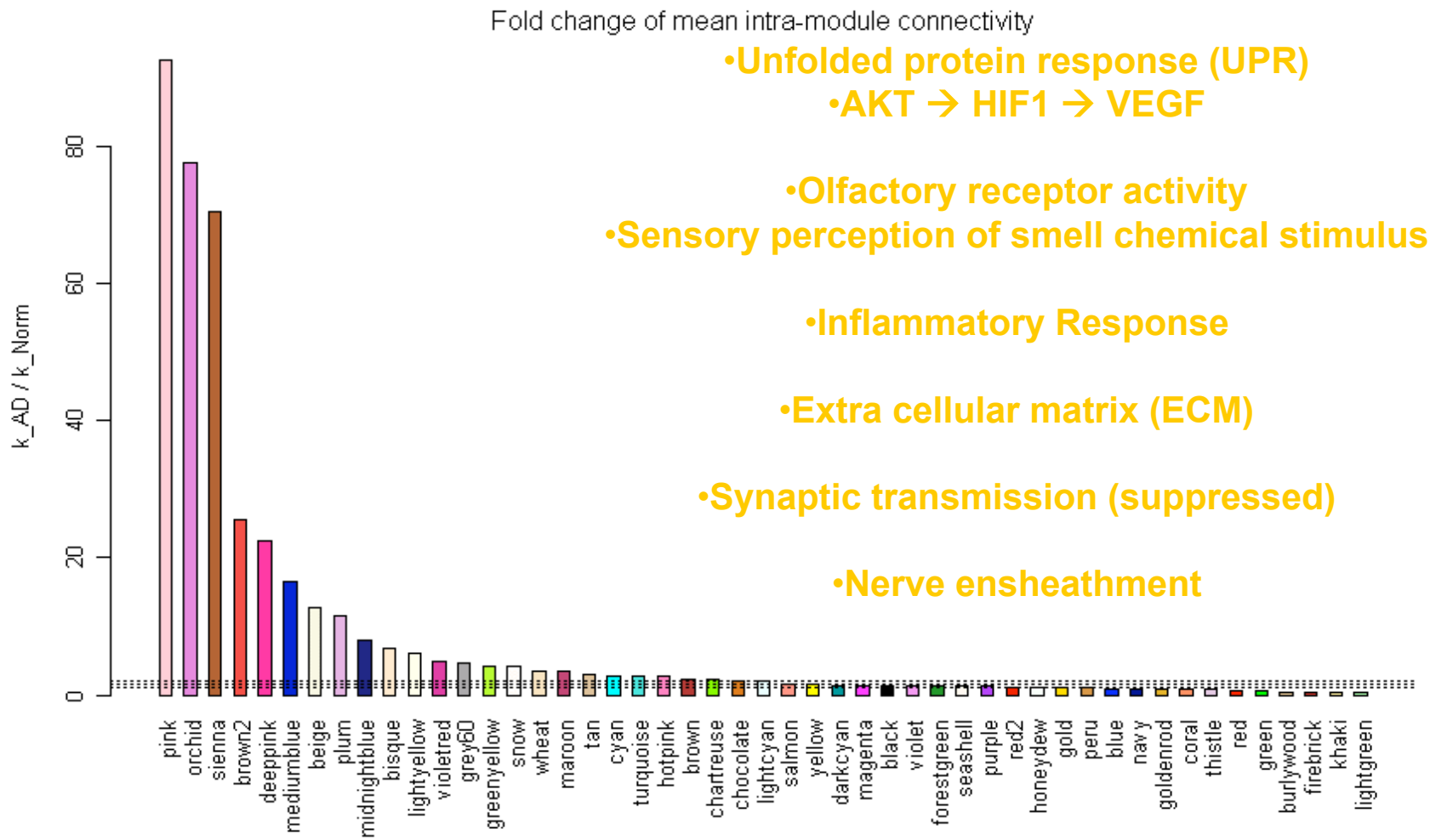
Gain connectivity by 1.9 fold





Module Connectivity Change (AD/Normal)

Differentially Connected Modules in AD



- **Unfolded protein response (UPR)**
- **AKT → HIF1 → VEGF**
- **Olfactory receptor activity**
- **Sensory perception of smell chemical stimulus**
- **Inflammatory Response**
- **Extra cellular matrix (ECM)**
- **Synaptic transmission (suppressed)**
- **Nerve ensheathment**

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

Sage Bionetworks

Global Coherent Data Sets

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

Sage Data Sets – Available/In transition

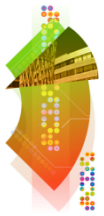
Available

Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
Mouse_CVD_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	CVD	Jake Lusis	UCLA	Sage - Available	334
Human_Cancer_HCC_HKU	HCC	Human	Cancer	John Luk	HKU	Sage - Available	250
Human_CVD_Liver_Vanderbilt/Pittsburgh/StJudes	Liver	Human	CVD	Guengrich/ Strom/ Schuetz	Vanderbilt/ Pittsburgh/ StJudes	Sage - Available	517

Transition

Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
Human_Cancer_Breast_BCCA	Breast	Human	Cancer	Aparicio/ Caldas	BCCA/ Cambridge	Sage - Available-	1,500
Human_Cancer_Glioblastoma_TCGA	Glioblastoma	Human	Cancer	TCGA	TCGA	Sage - Available-subset	413
Human_Neurodegenerative_Brain:Prefrontal cortex_Visual Cortex_Cerbellum_HBTRC	Brain:Prefrontal cortex_Visual Cortex_Cerbellum	Human	Neuro-degenerative	Francine Benes	HBTRC	Sage - Transition	700
Mouse_Metabolic_Liver_UCLA	Liver	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	111
Human_Cancer_AML(pediatric)_FHCRC	AML(pediatric)	Human	Cancer	Sohail Meschinchi	FHCRC	Sage - Transition	200
Mouse_Metabolic_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	442
Mouse_Metabolic_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	309

Note that for Datasets with >1 tissue the "approximate numbers of individuals" refers to the tissue with the greatest number of individuals and all tissues may not have this degree of coverage



Sage Data Sets – Requires Release

MGED

Requires Release

Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
Human_Cancer_Breast_Stanford	Breast	Human	Cancer	Jonathan Pollack	Stanford	Requires Release	89
Human_Cancer_Breast_KI	Breast	Human	Cancer	Jonas Bergh	Karolinska	Requires Release	650
Human_Metabolic_Adipose_Blood_deCODE	Adipose_Blood	Human	Metabolic	Kari Stefansson	deCODE	Requires Release	1,002
Human_Metabolic_Adipose:Omental/subQ_Liver_Stomach_Harvard	Adipose:Omental/subQ_Liver_Stomach	Human	Metabolic	Lee Kaplan	Harvard	Requires Release	975
Human_Respiratory_Lung_UBC/Laval/Groningen	Lung	Human	Respiratory	Parre/Bosse/Timmens	UBC/Laval/Groningen	Requires Release	1,180
Human_Metabolic_Adipose_Muscle_NIDDK	Adipose_Muscle	Human	Metabolic	Cliff Bogardus	NIDDK	Requires Release	225
Human_Multiple_Blood_NHLBI	Blood	Human	Multiple	Dan Levy	NHLBI	Requires Release	5,000
Human_CVD_Macrophage_Liver_Carotid_IMA_Adipose_Muscle_Plaque_Blood_Karolinska	Macrophage_Liver_Carotid_IMA_Adipose_Muscle_Plaque_Blood	Human	CVD	Johan Bjorkegren	Karolinska	Requires Release	100
Human_Inflammatory Disease_Blood_UCL	Blood	Human	Inflammatory Disease	David van Heel	UCL	Requires Release	1,469
Human_Asthma_Blood_Imperial	Blood	Human	Asthma	William Cookson	Imperial	Requires Release	400
Human_CVD_PBMC_UTSW	PBMC	Human	CVD	John Blangero	UTSW	Requires Release	1,240
Mouse_Metabolic_Islet_Liver_Adipose_Hypothalamus_Kidney_Muscle_Wisconsin	Islet_Liver_Adipose_Hypothalamus_Kidney_Muscle	Mouse	Metabolic	Alan Attie	Wisconsin	Requires Release	500
Mouse_Sleep_Hypothalamus_Thalamus_Frontal Cortex_Liver_Northwestern	Hypothalamus_Thalamus_Frontal Cortex_Liver	Mouse	Sleep	Fred Turek	Northwestern	Requires Release	250
Mouse_Sleep_Hypothalamus_Thalamus_Frontal Cortex_Hippocampus_Northwestern	Hypothalamus_Thalamus_Frontal Cortex_Hippocampus	Mouse	Sleep	Fred Turek	Northwestern	Requires Release	220
Mouse_Metabolic_Adipose_Liver_Duodenum_Islets_UCLA	Adipose_Liver_Duodenum_Islets	Mouse	Metabolic	Jake Lusis	UCLA	Requires Release	500
Mouse_Respiratory_Lung_Harvard	Lung	Mouse	Respiratory	David Beier	Harvard	Requires Release	200
Mouse_Metabolic_Liver_Adipose_Hypothalamus_Arizona	Liver_Adipose_Hypothalamus	Mouse	Metabolic	Daniel Pomp	Arizona	Requires Release	308
Mouse_Metabolic_Adipose_Liver_Hypothalamus_Muscle_Merck	Adipose_Liver_Hypothalamus_Muscle	Mouse	Metabolic	Eric Schadt	Merck	Requires Release	1,650
Mouse_CVD_Adipose_Kidney cortex_Kidney medulla_Liver_Merck	Adipose_Kidney cortex_Kidney medulla_Liver	Mouse	CVD	Eric Schadt	Merck	Requires Release	350
Mouse_Multiple_Multiple Tissues_Tennessee	Multiple Tissues	Mouse	Multiple	Rob Williams	Tennessee	Requires Release	60
Mouse_CVD_Liver_UCLA	Liver	Mouse	CVD	Jake Lusis	UCLA	Requires Release	100
Mouse_Multiple_Hippocampus_Lung_Liver_Oxford	Hippocampus_Lung_Liver	Mouse	Multiple	Jonathan Flint	Oxford	Requires Release	1,900
Mouse_Sarcopenia_Muscle_PSU	Muscle	Mouse	Sarcopenia	Arimantas Lionikas	PSU	Requires Release	811
Mouse_Cancer_Skin_UCSF	Skin	Mouse	Cancer	Allan Balmain	UCSF	Requires Release	71
Human_Cancer_Lung_Mayo	Lung	Human	Cancer	Ping Yang	Mayo	Requires Release	70
Mouse_Cancer_Breast_UNC	Breast	Mouse	Cancer	Ryan Gordon	UNC	Requires Release	615
Human_Alzheimer_Cortex_Miami	Cortex	Human	Alzheimer	Amanda Myers	Miami	Requires Release	364

Sage Data Sets – In Progress

In progress

Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
Human_Cancer_Ovarian_TCGA	Ovarian	Human	Cancer	TCGA	TCGA	In progress	387
Human_Cancer_Lung_TCGA	Lung	Human	Cancer	TCGA	TCGA	In progress	500
Human_Cancer_HCC_NCI	HCC	Human	Cancer	Snorri Thorgeirsson	NCI	In progress	200
Human_Cancer_Lung_Canary	Lung	Human	Cancer		Canary	In progress	
Human_Cancer_Pancreatic_ICGC	Pancreatic	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Ovarian_ICGC	Ovarian	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Gastric_ICGC	Gastric	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Breast(triple negative)_ICGC	Breast(triple negative)	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Breast(HER2)_ICGC	Breast(HER2)	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_HCC (viral)_ICGC	HCC (viral)	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_HCC (alcohol)_ICGC	HCC (alcohol)	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Brain(pediatric)_ICGC	Brain(pediatric)	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Oral_ICGC	Oral	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_CLL_ICGC	CLL	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Glioblastoma_ICGC	Glioblastoma	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Lung_ICGC	Lung	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_AML_ICGC	AML	Human	Cancer	ICGC	ICGC	In progress	500
Human_Cancer_Colon_ICGC	Colon	Human	Cancer	ICGC	ICGC	in progress	500
Human_Cancer_Lung_ACRG	Lung	Human	Cancer	ACRG	ACRG	In progress	2,000
Human_Cancer_Gastric_ACRG	Gastric	Human	Cancer	ACRG	ACRG	In progress	2,000
Human_Cancer_Myeloma_Sage	Myeloma	Human	Cancer	Stephen Friend	Sage	In progress	300
Human_Cancer_Lung_Sage	Lung	Human	Cancer	Stephen Friend	Sage	In progress	300
Human_Cancer_Ovarian_Sage	Ovarian	Human	Cancer	Stephen Friend	Sage	In progress	300
Human_Cancer_AML_Sage	AML	Human	Cancer	Stephen Friend	Sage	In progress	300
Human_Cancer_Breast_Sage	Breast	Human	Cancer	Stephen Friend	Sage	In progress	300
Human_Cancer_Medulloblastoma_JHSM	Medulloblastoma	Human	Cancer	Bert Vogelstein	JHSM	In progress	47
Human_Cancer_Pancreas_JHSM	Pancreas	Human	Cancer	Bert Vogelstein	JHSM	In progress	47
Human_Cancer_Breast_NKI	Breast	Human	Cancer	Rene Bernards	NKI	In progress	1,000
Human_Cancer_Colon_HKU	Colon	Human	Cancer	Suet-Yi Leung	HKU	In progress	400

Types of Network 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/StJudes	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

Sage 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

Sage BIONETWORKS

Bioconductor at Sage

- Bioconductor standard part of data QC SOP
  - Use `Biobase` and `affy` to read and store data
  - Use same for normalization and transformation
  - Use `genefilter` for subset selection
- Use Bioconductor for mouse data analysis
  - Use `qtl` package for analysis of F_2 mouse cross
  - Use custom scripts for report plots and figures

New Packages at Sage- Dave Henderson

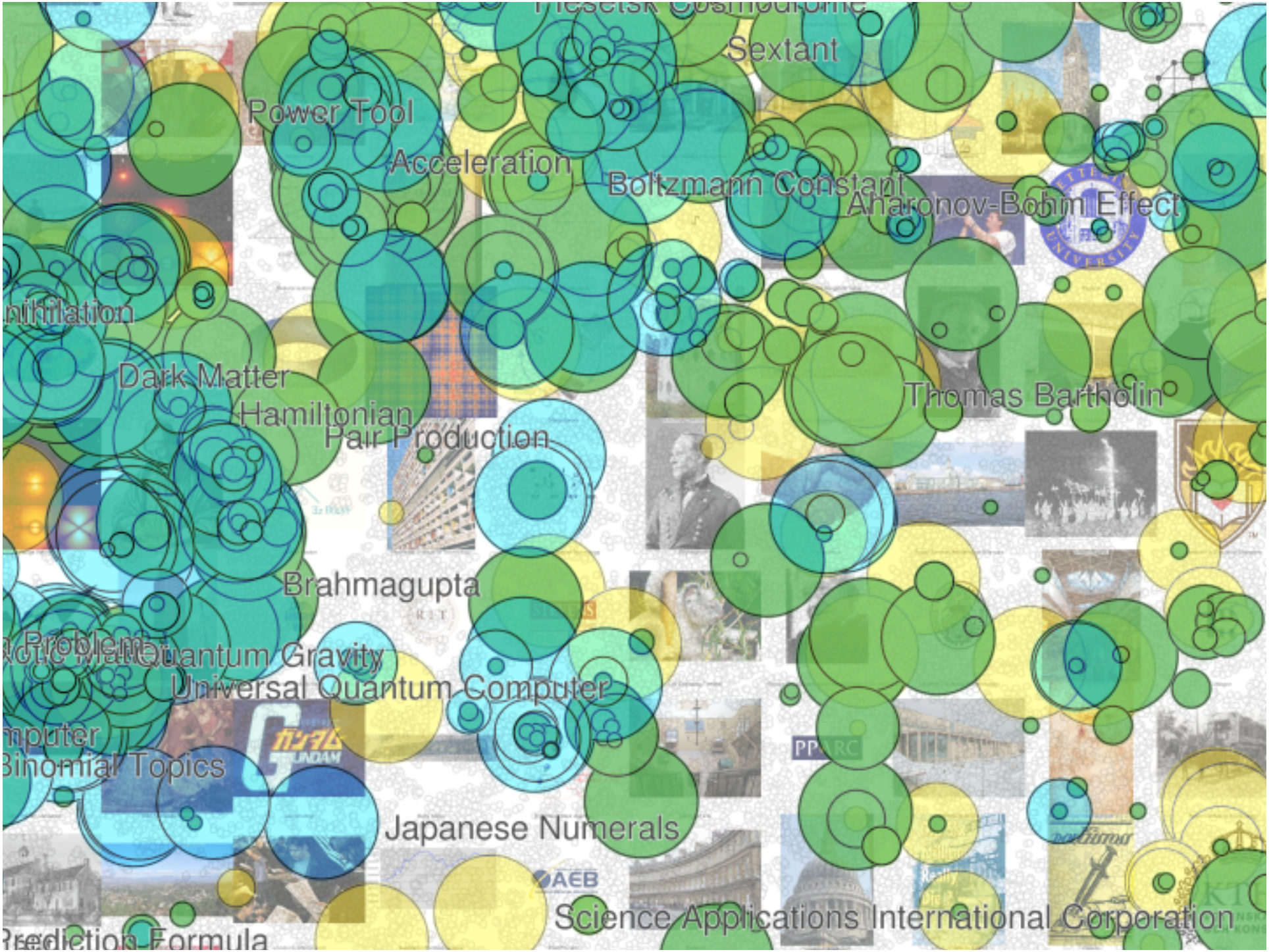
- Causal Inference Test
  - R package under development at Sage
  - Performs a test for causal inference developed by Joshua Millstein
- Key Driver
  - R package under development at Sage
  - Identifies key nodes within a graph given a query set of nodes
- Sage Data Sets
  - Contains the publicly available data sets from Sage
  - Data in native Bioconductor objects
  - GraphSet object under development with Bioconductor group in Seattle

NOT JUST WHAT BUT HOW

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

attribution: carole goble- sidney brenner


this must be accessible and be integrated


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

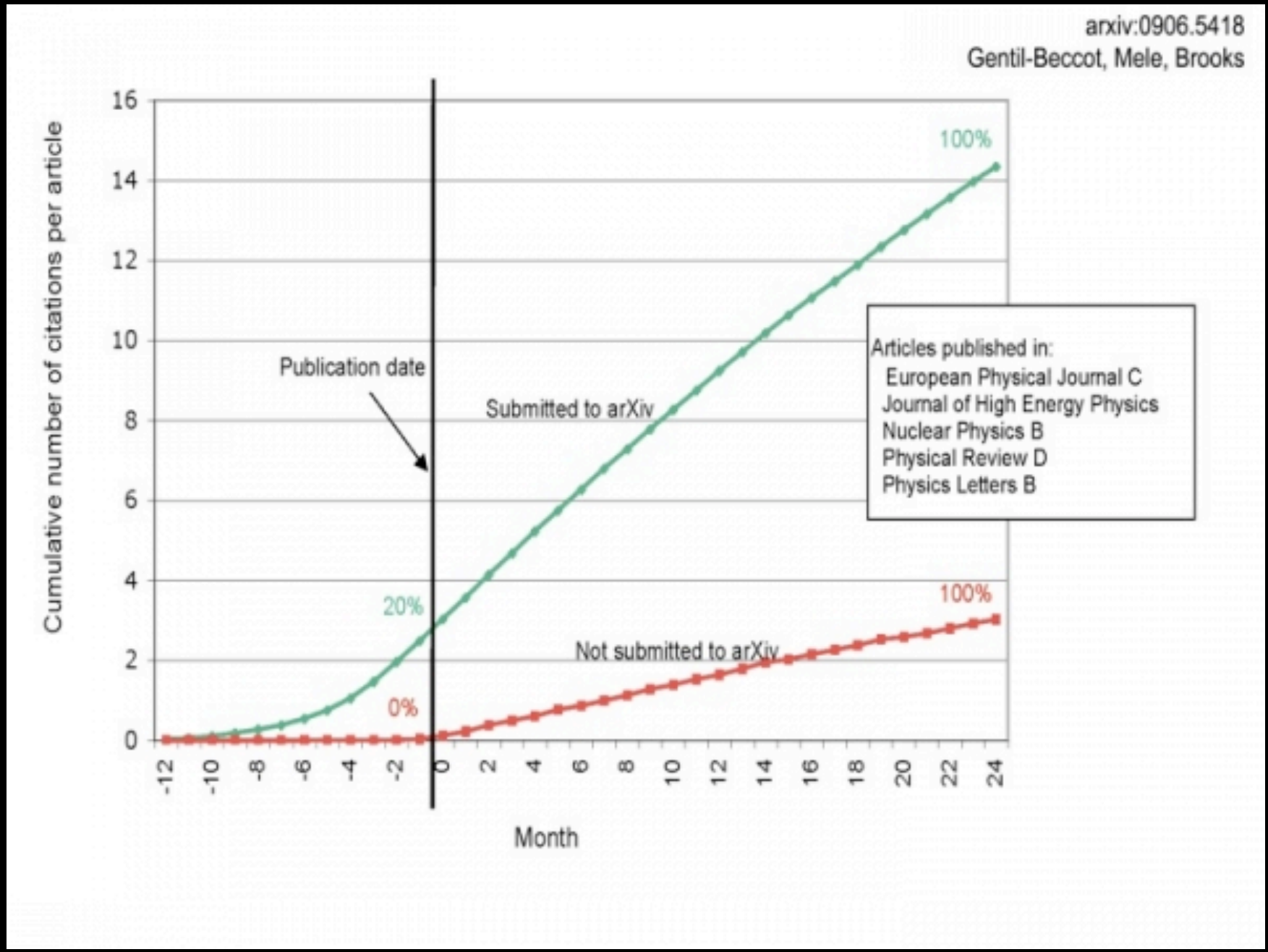
Open Access Use in Physics

Gentil-Beccot, Anne; Salvatore Mele, Travis Brooks (2009) Citing and Reading Behaviours in High-Energy Physics: How a Community Stopped Worrying about Journals and Learned to Love RepositoriesThis is an important study, and most of its conclusions are valid:

(1) Making research papers open access (OA) dramatically increases their impact

(2) .(2) The earlier that papers are made OA, the greater their impact

(3) .(3) High Energy Physics (HEP) researchers were among the first to make their papers OA (since 1991, and they did it without needing to be mandated to do it!)


arxiv:0906.5418
Gentil-Beccot, Mele, Brooks
Cumulative number of citations per article
Month
Publication date
Submitted to arXiv
Not submitted to arXiv
100%
20%
100%
0%
Articles published in:
European Physical Journal C
Journal of High Energy Physics
Nuclear Physics B
Physical Review D
Physics Letters B

SGC-Increasing accessibility of Structural Biology data

- Oct 09: Collaborate with PLoS ONE to produce PLoS ONE collection

- Aimed at non-structural biologists
- Peer reviewed
- Published in a novel electronic annotated format, incorporate 3D visualisation
  - Use MolSoft LLC activeICM technology

MolSoft molecules *in silico*

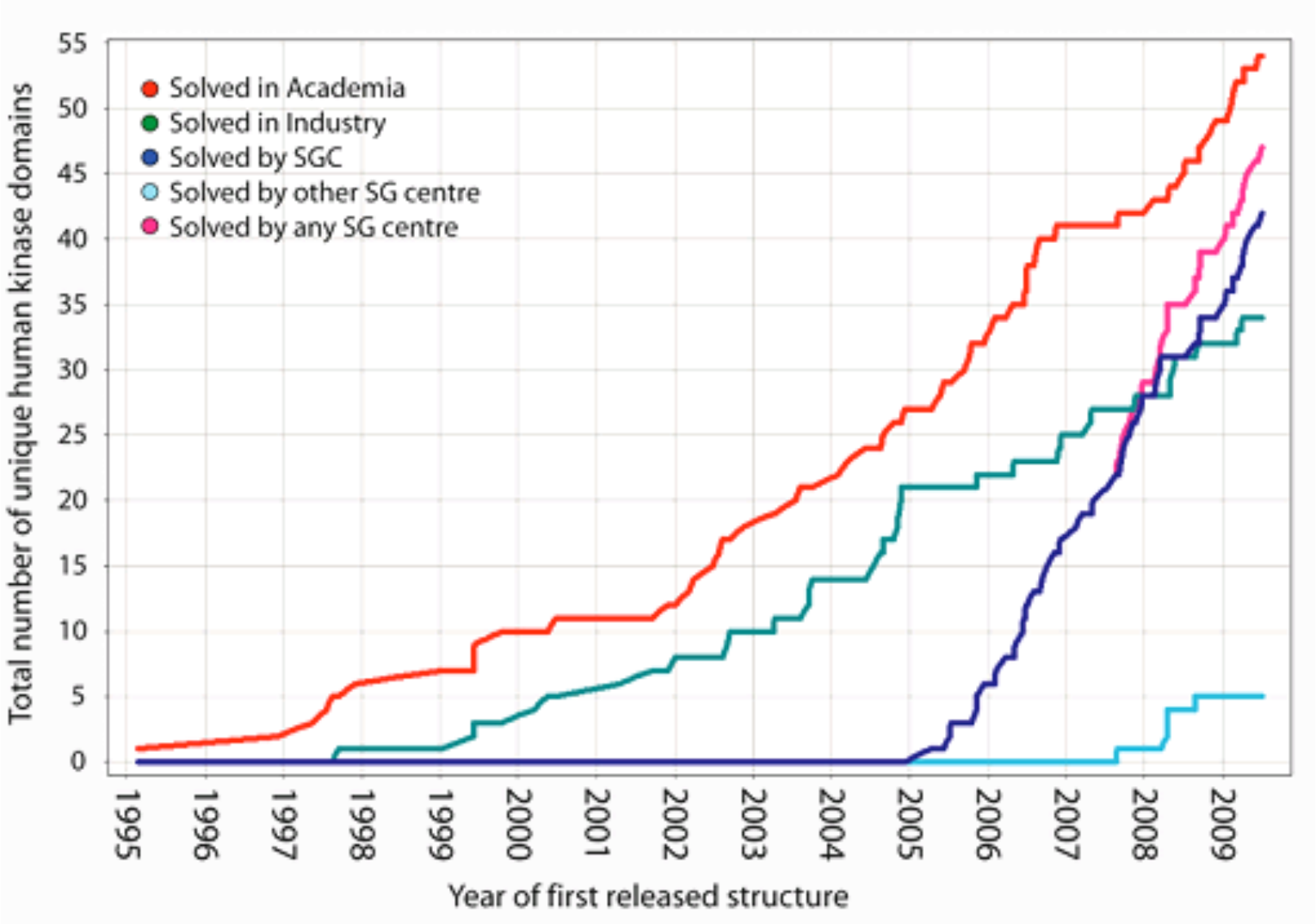
iSee

PLoS one

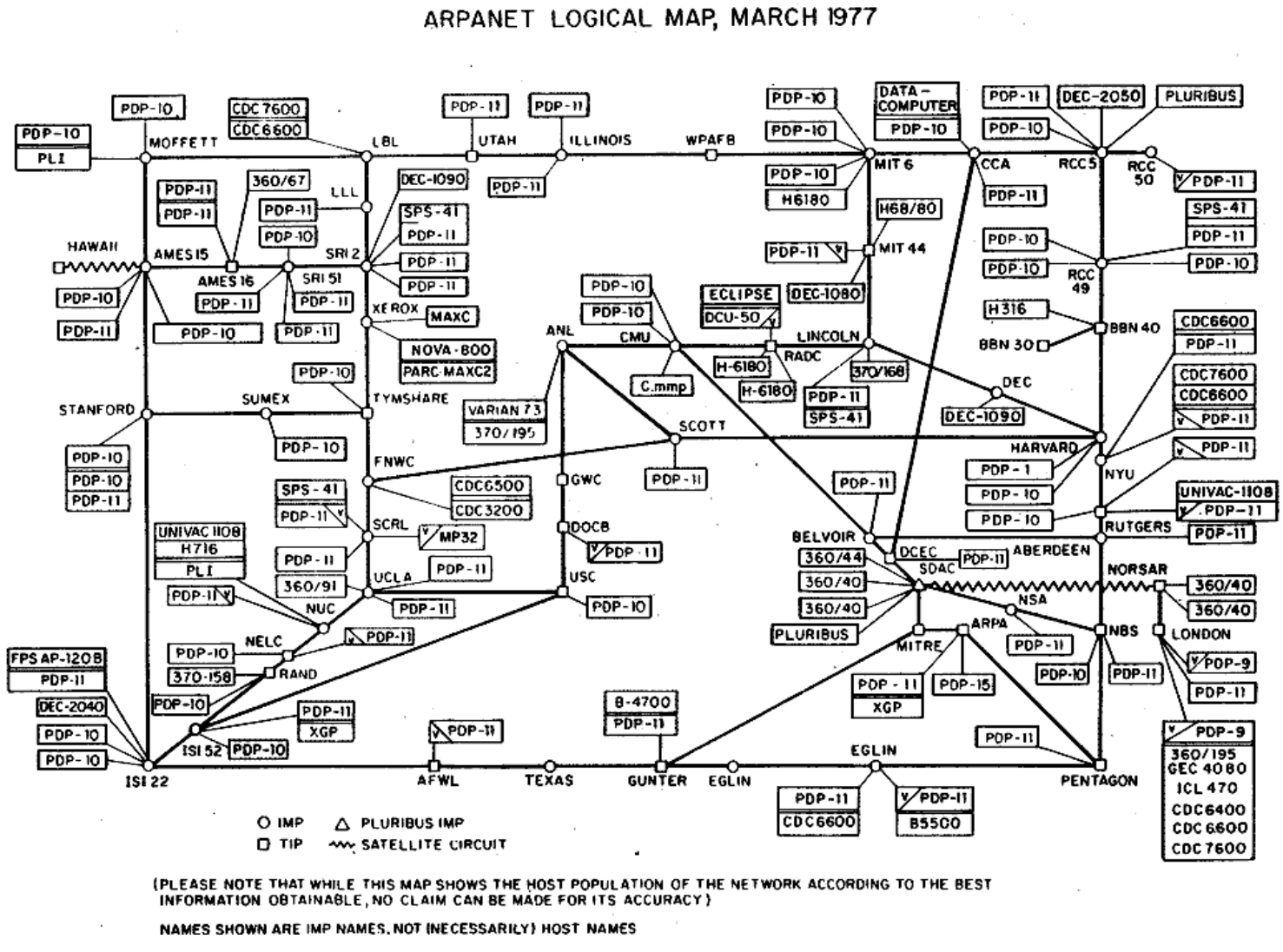
Impact of “no IP”

- Collaborate quickly with any scientist, lab or institution
- Work closely with multiple private organisations, on same project
- Generate data quickly
- Place data in public domain quickly

40% of all kinase structures, solved by SGC (in past 4 years



BUILDING A COMMONS FOR EVOLVING GENERATIVE MODELS OF DISEASE



(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

Sage BIONETWORKS

Home | Company | Research | Commons | Training

Home › Commons

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.

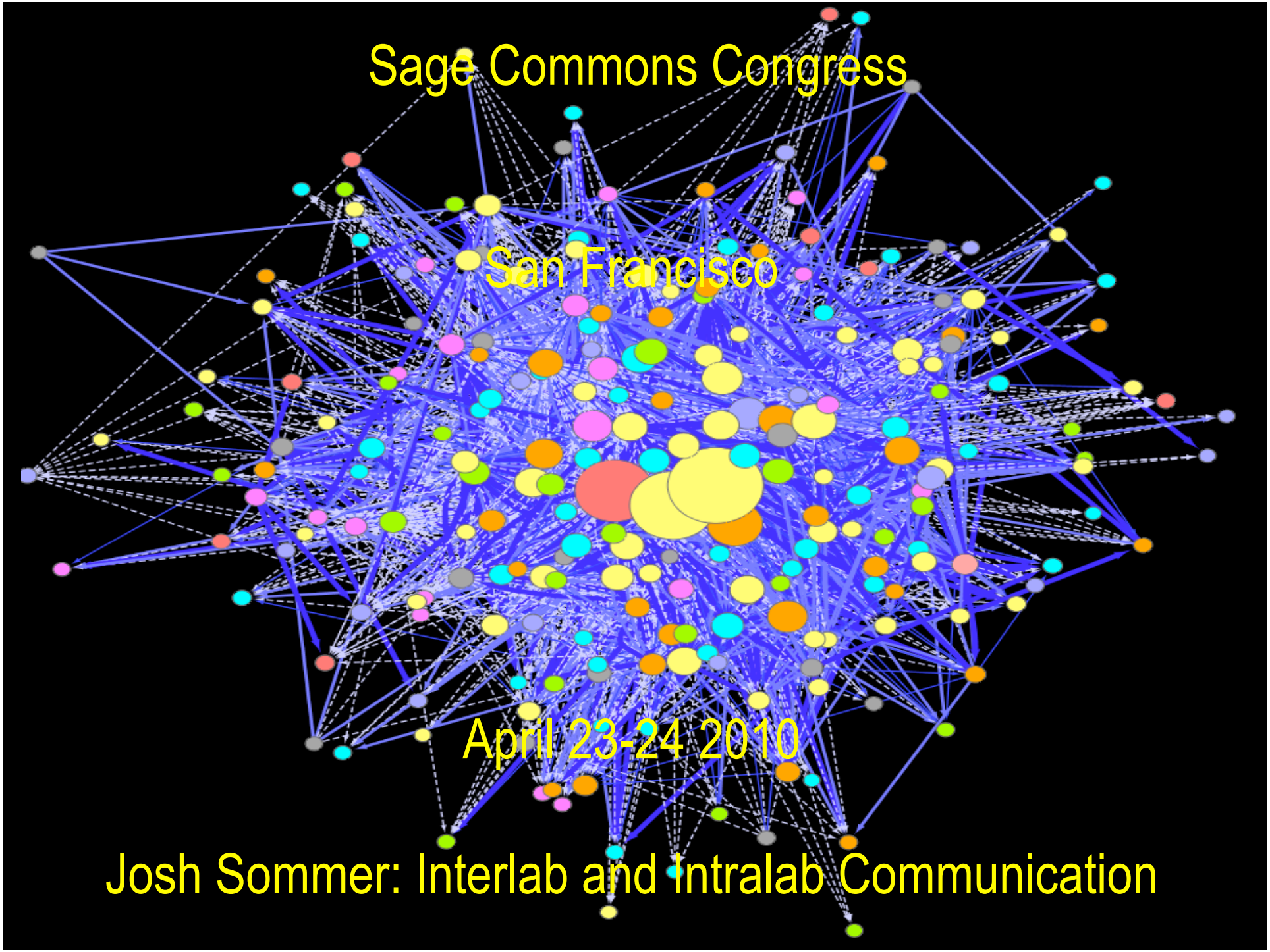
Integration of molecular mega-data sets

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 Sommer: 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

Opportunities to Leverage Existing Efforts

Bioconductor

caBIG

Cytoscape

Genome Space

Wikipathways

BETTER MAPS OF DISEASE

NOT JUST WHAT BUT HOW

BUILDING A COMMONS FOR EVOLVING GENERATIVE MODELS OF DISEASE

Current big science efforts to interpret the biology thru DNA changes, RNA changes, proteomic changes layered on existing signal and metabolic pathways represent fragmented approaches

The stunning technologies coming will generate heaps of data that is expanding faster than we can process it.

Current biomedical approaches to developing therapies starting from RO1 driven academic labs to existing pharma/biotechs collecting siloed insights driving existing clinical approvals are unsustainable

Emerging efforts to build bionetworks using integrative genomic approaches can highlight the selected components in diseases that are non-redundant, (minimal redundancy) and therefore if changes can produce be drivers of the disease and drivers of therapies

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