Integrative *causal* networks for understanding complex human diseases

Jun Zhu, Ph. D.

Professor of Genomics and Genetic Sciences

Icahn Institute of Genomics and Multi-scale Biology

The Tisch Cancer Institute

Icahn Medical School at Mount Sinai

New York, NY

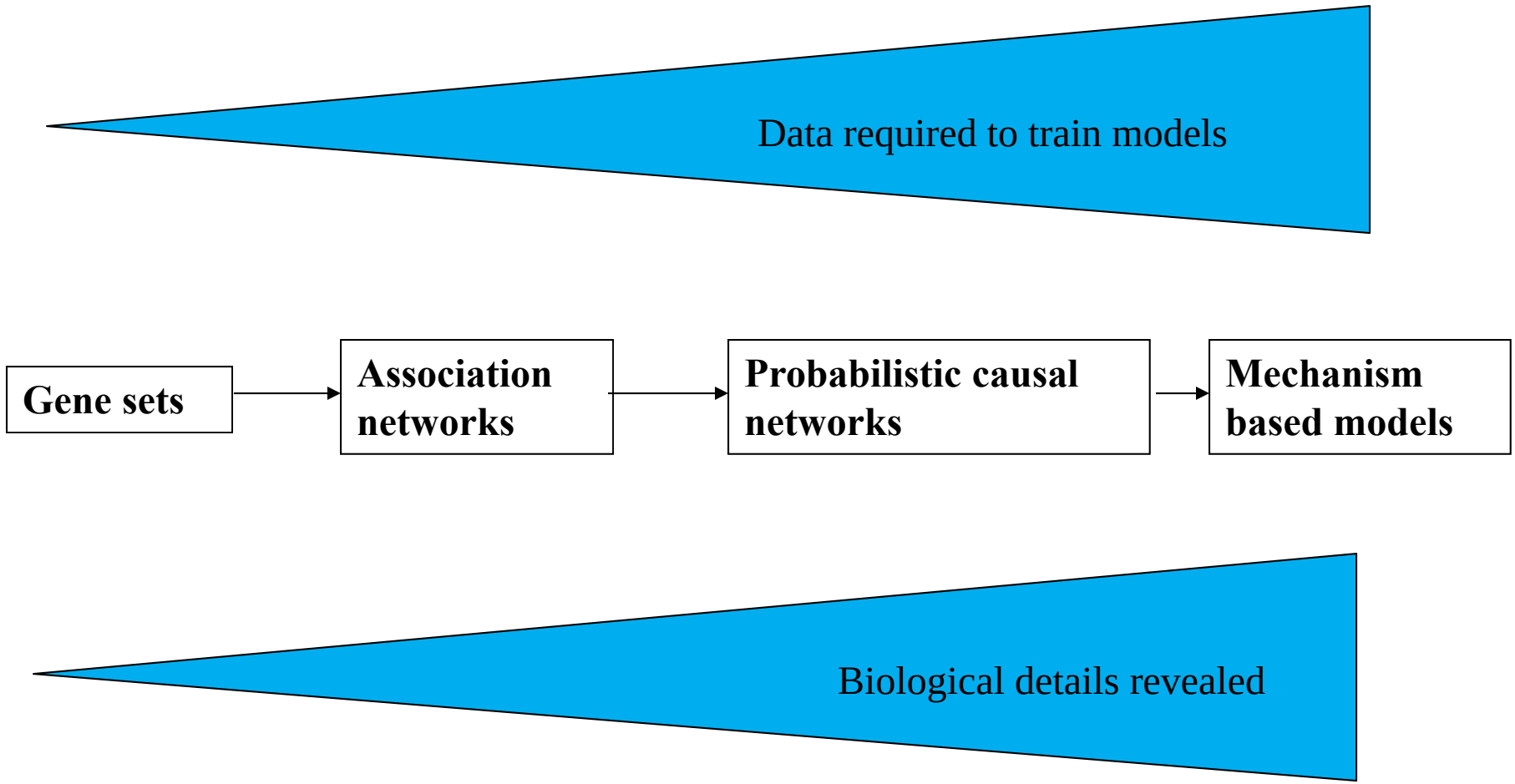
http://research.mssm.edu/integrative-network-biology/

Email: jun.zhu@mssm.edu

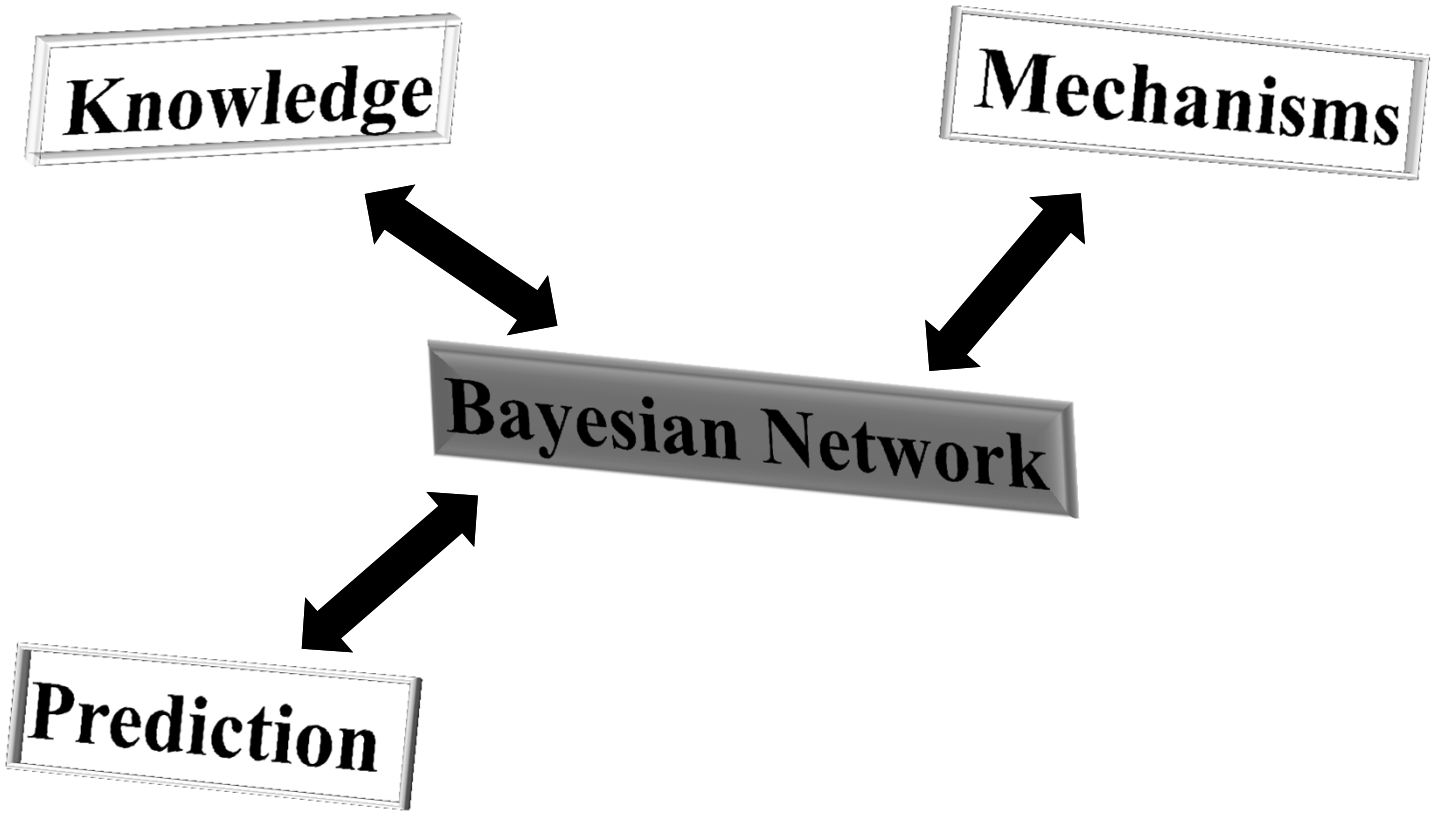
@IcahnInstitute

Pittsburg University, October, 2015

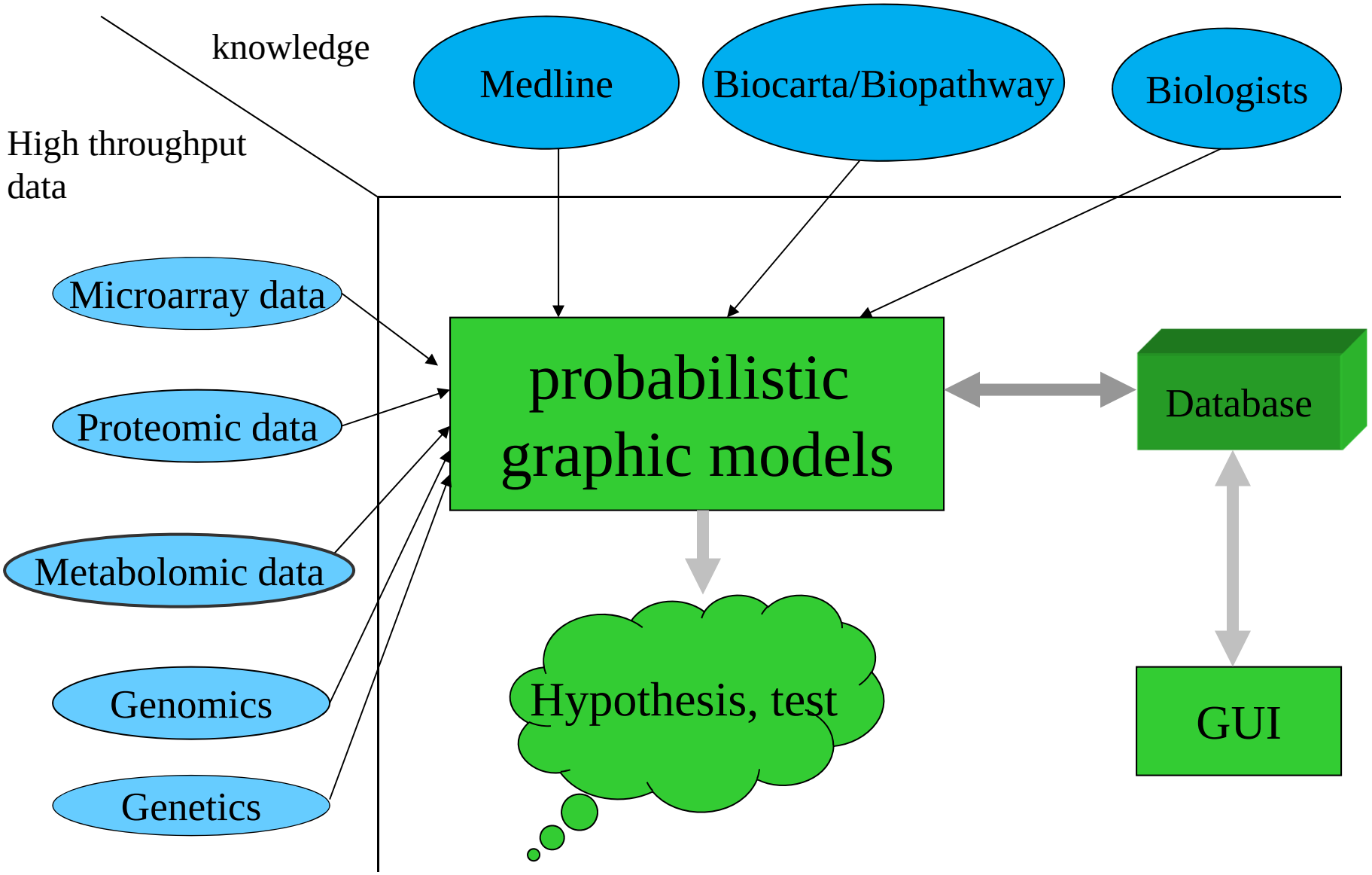
Biological networks/pathways



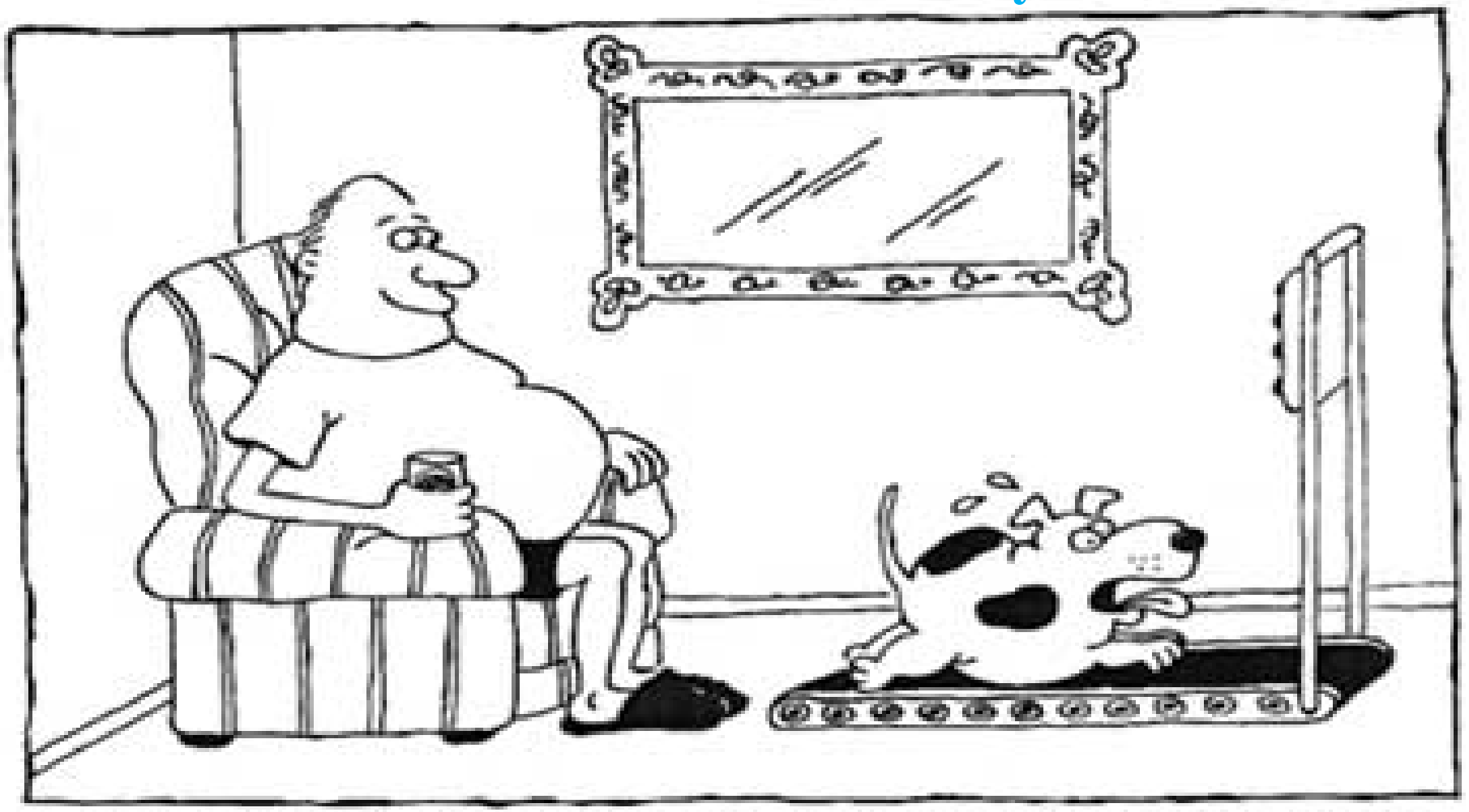
Why Bayesian network is chosen for causal modelling?



A framework for building biological causal networks



Association vs Causality



WHEN INFORMED BY HIS DOCTOR OF THE CORRELATION BETWEEN FAT DOGS AND THEIR MASTERS, BRIAN SET OUT IMMEDIATELY TO RECTIFY HIS WEIGHT PROBLEM.

From Stephen Friend

atherosclerosis 1996

High-density lipoprotein cholesterol as a predictor of coronary heart disease risk. The PROCAM experience and pathophysiological implications for reverse cholesterol transport

Gerd Assmann, Helmut Schulte, Arnold von Eckardstein, Yadong Huang

Abstract

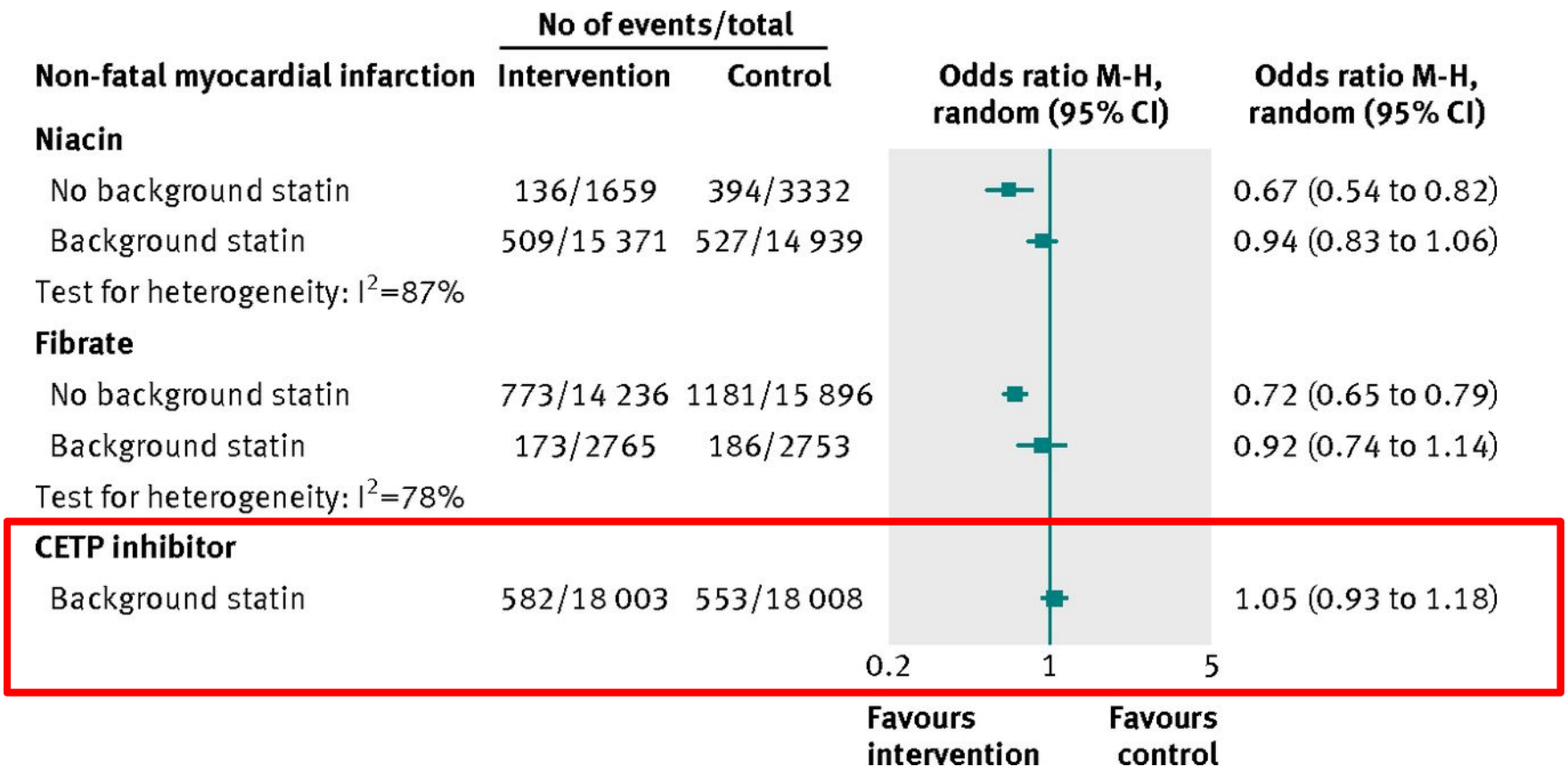
The incidence of coronary heart disease (CHD) was assessed via the Prospective Cardiovascular Münster (PROCAM) study in 19 698 volunteer subjects aged between 16 and 65 years. An adequate incidence of atherosclerotic CHD was only found in male subjects greater than 40 years of age. The analysis and subsequent 6 year follow-up period was, therefore, confined to 4559 male participants aged 40–64 years. In the follow-up period, 186 study participants developed atherosclerotic CHD (134 definite non-fatal myocardial infarctions (MIs) and 52 definite atherosclerotic CHD deaths including 21 sudden cardiac deaths and 31 fatal MIs). Univariate analysis revealed a significant association between the incidence of atherosclerotic CHD and high-density lipoprotein cholesterol ($P < 0.001$), which remained after adjustment for other risk factors.

Table 3| **Description of included studies of cholesteryl ester transfer protein (CETP) inhibitors**

Reference	Trial drugs and dose	Control	Follow-up (months)	No enrolled (No intervention, No control)	Statin use (%)	Men (No intervention, No control)	Mean (SD) age (years) (intervention, control)	White ethnicity (%)	Increase in HDL from baseline in active arm (%)
Dal-OUTCOMES[33] 2012	Dalcetrapib 600 mg daily	Placebo	31	15 871 (7938, 7933)	97	6365, 6436	60.3 (9.1), 60.1 (9.1)	88	40
Dal-PLAQUE[34] 2011	Dalcetrapib 600 mg daily	Placebo	24	130 (64, 66)	87	51, 55	62.6 (8.2), 64.6 (7.8)	92	31
Dal-VESSEL[35] 2012	Dalcetrapib 600 mg daily	Placebo	8	476 (239, 237)	95	211, 211	62.3 (7.05), 61.9 (7.92)	NR	31
Define[36] 2010	Anacetrapib 100 mg daily	Placebo	18	1623 (811, 812)	99	629, 618	62.5 (8.7), 62.9 (9.0)	83	138
Illuminate[37] 2007	Torcetrapib 60 mg daily	Placebo	18	15 054 (7528, 7526)	100	5854, 5861	61.3 (7.6), 61.3 (7.6)	93	72
Illustrate[38] 2007	Torcetrapib 60 mg daily	Placebo	24	1188 (591, 597)	100	416, 421	56.9 (9.1), 57 (9.2)	NR	61
Radiance 1[39] 2007	Torcetrapib 60 mg daily	Placebo	24	850 (423, 427)	100	214, 232	46.8 (12.0), 45.2 (12.9)	NR	52
Radiance 2[40] 2007	Torcetrapib 60 mg daily	Placebo	20	752 (377, 375)	100	237, 245	57.9 (8.1), 56.5 (8.2)	NR	63

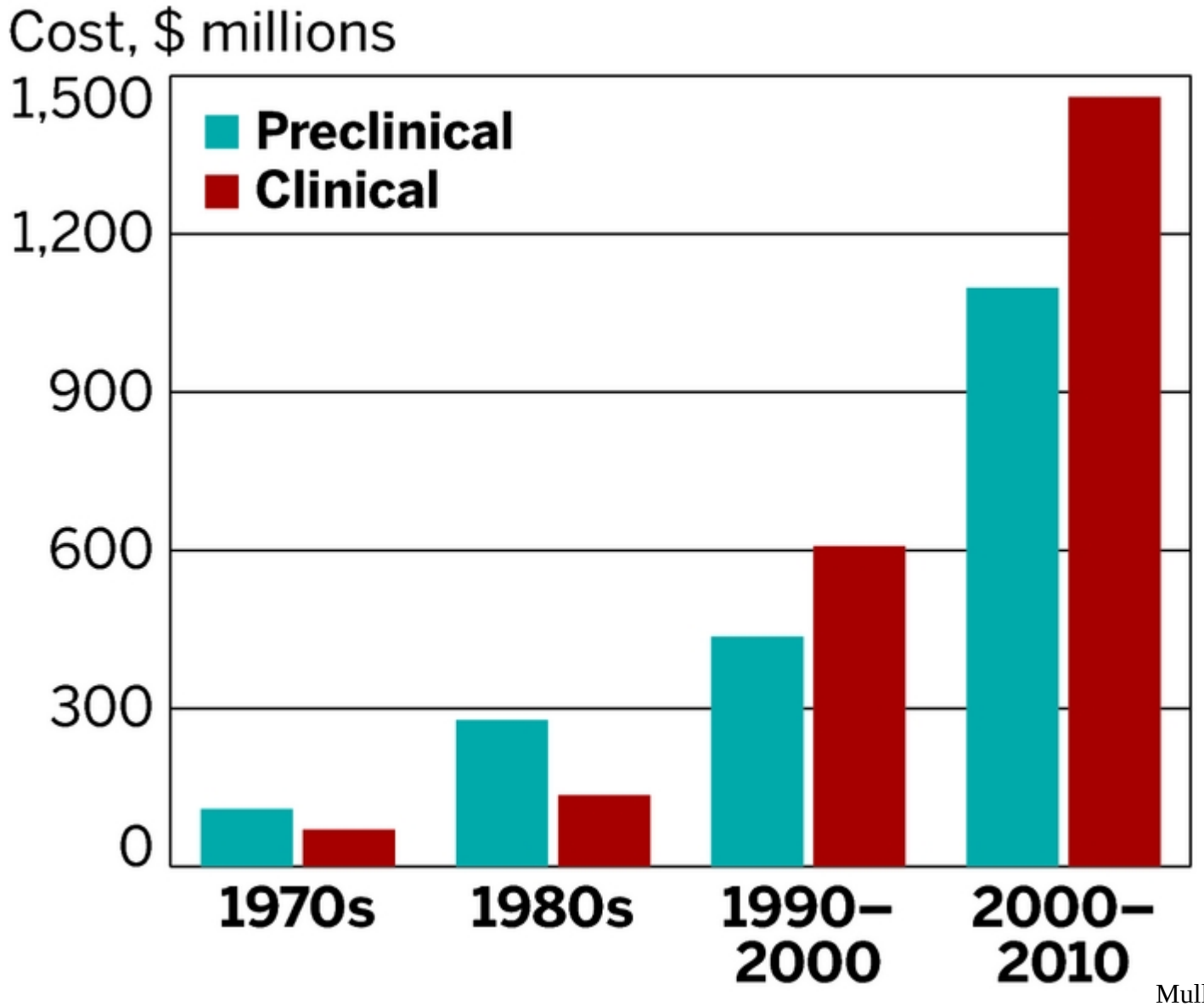
HDL=high density lipoprotein; NR=not reported.

Fig 4 The statin revolution: without background statin treatment, fibrates and niacin were found to reduce non-fatal myocardial infarction.



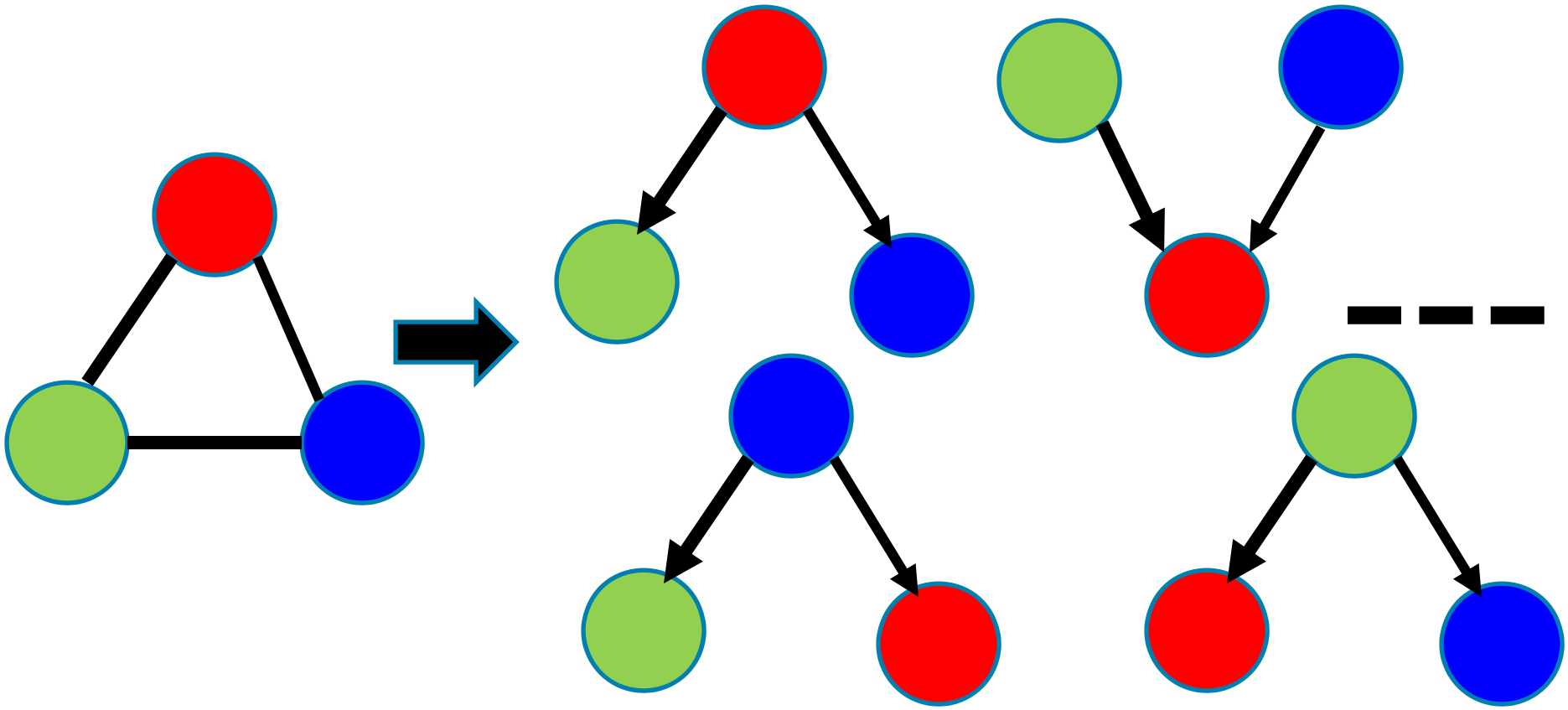
Daniel Keene et al. BMJ 2014;349:bmj.g4379

The cost of developing a prescription drug that gains market approval

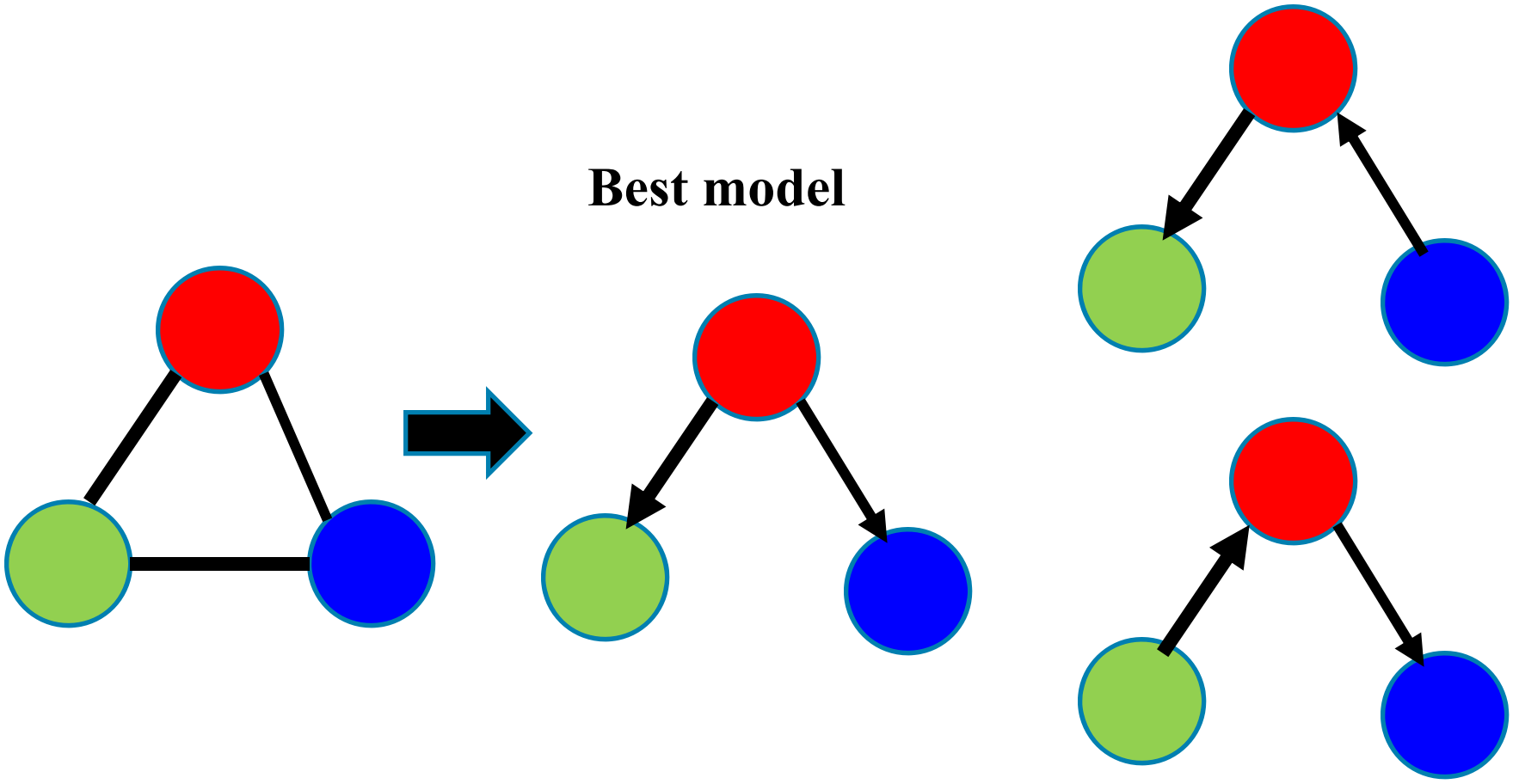


Mullin Scie. Ameri. 2014

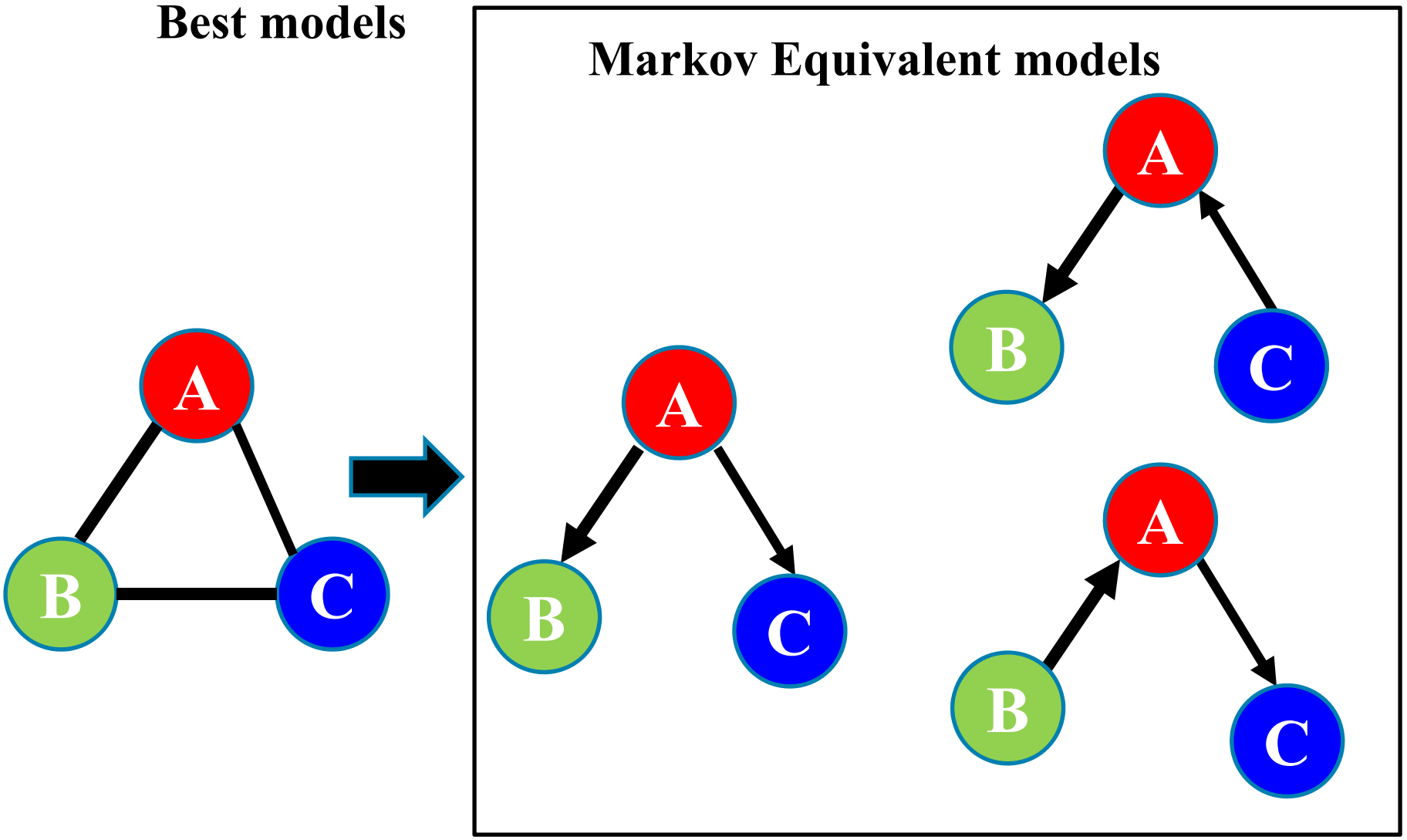
A simple biological question: are there causal/reactive relationships?



A Bayesian network approach:

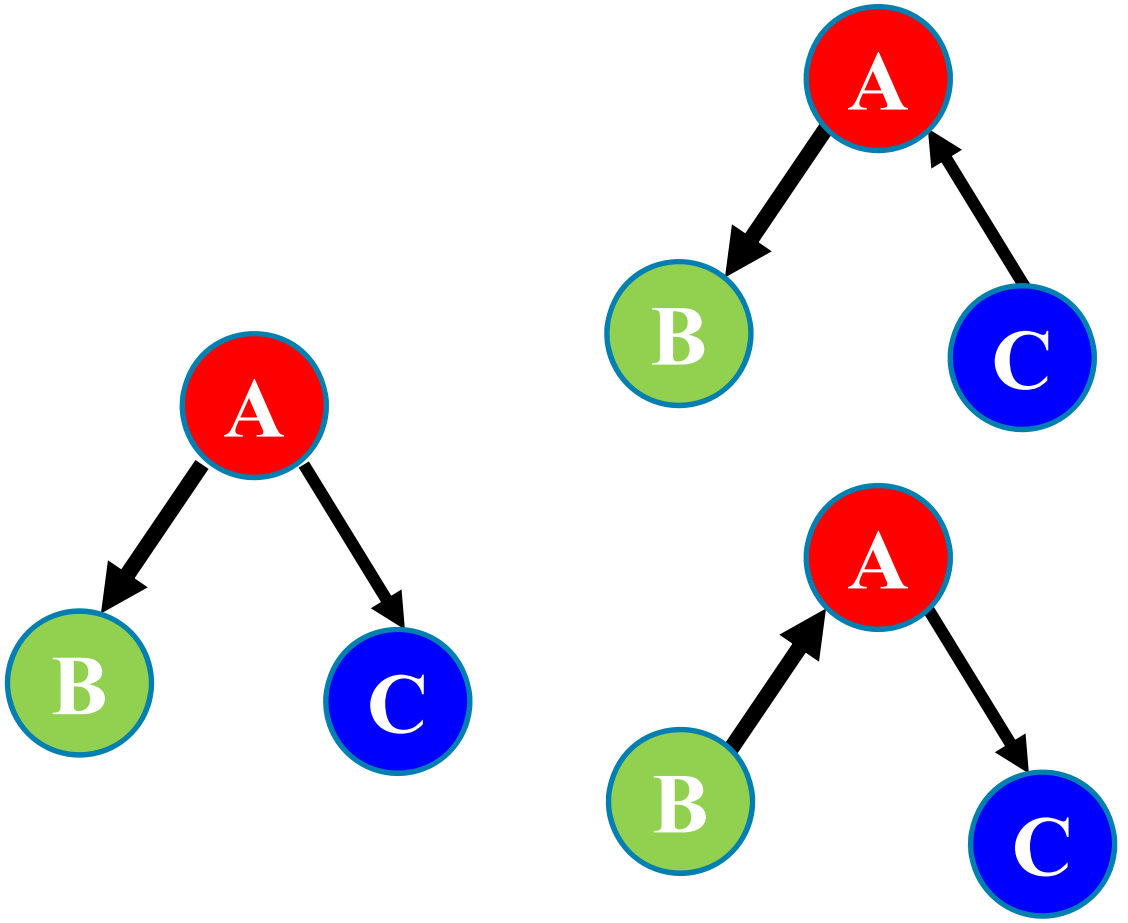


A Bayesian network approach:



A Bayesian network ≠ a causal structure

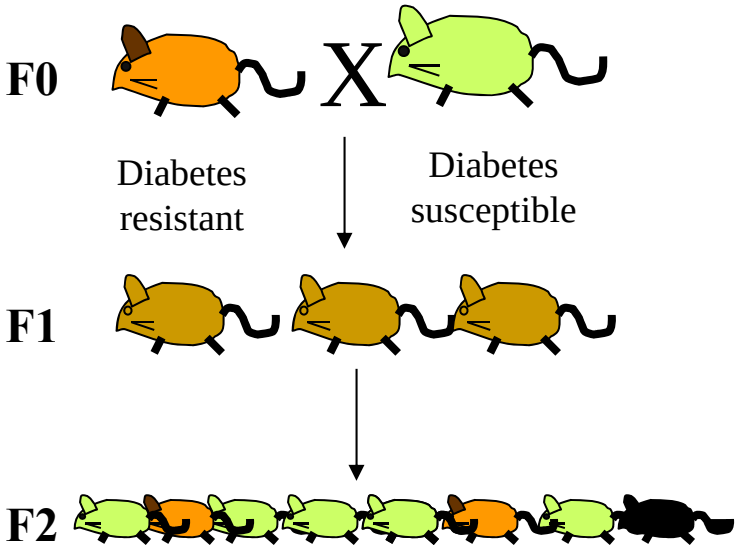
Markov Equivalent models

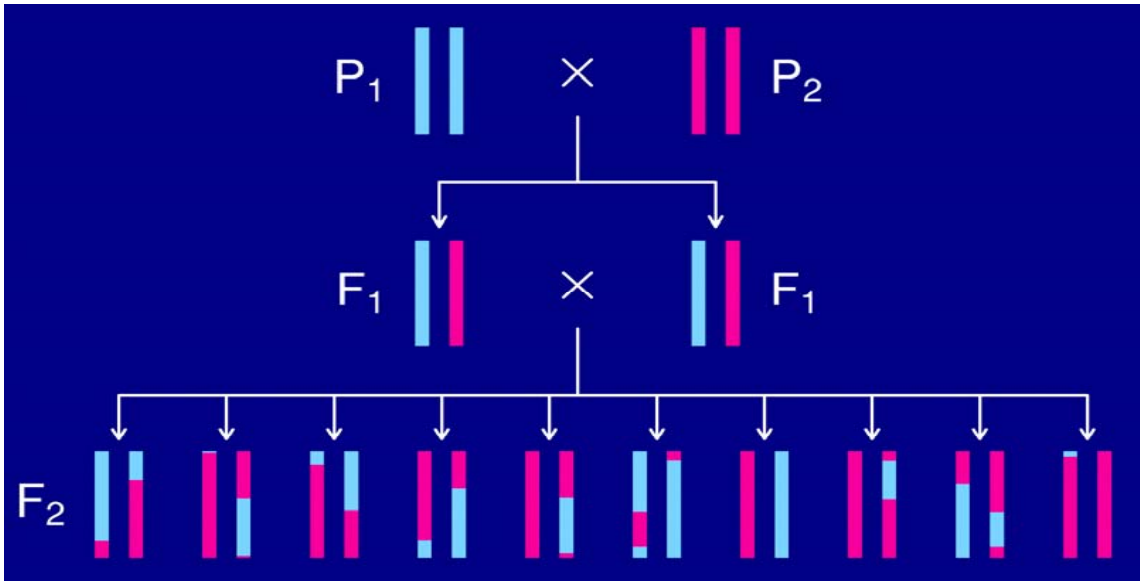


$$B \perp C \mid A$$

Bayesian network: how to break Markov equivalent?

Animal model: mouse F2 intercrosses

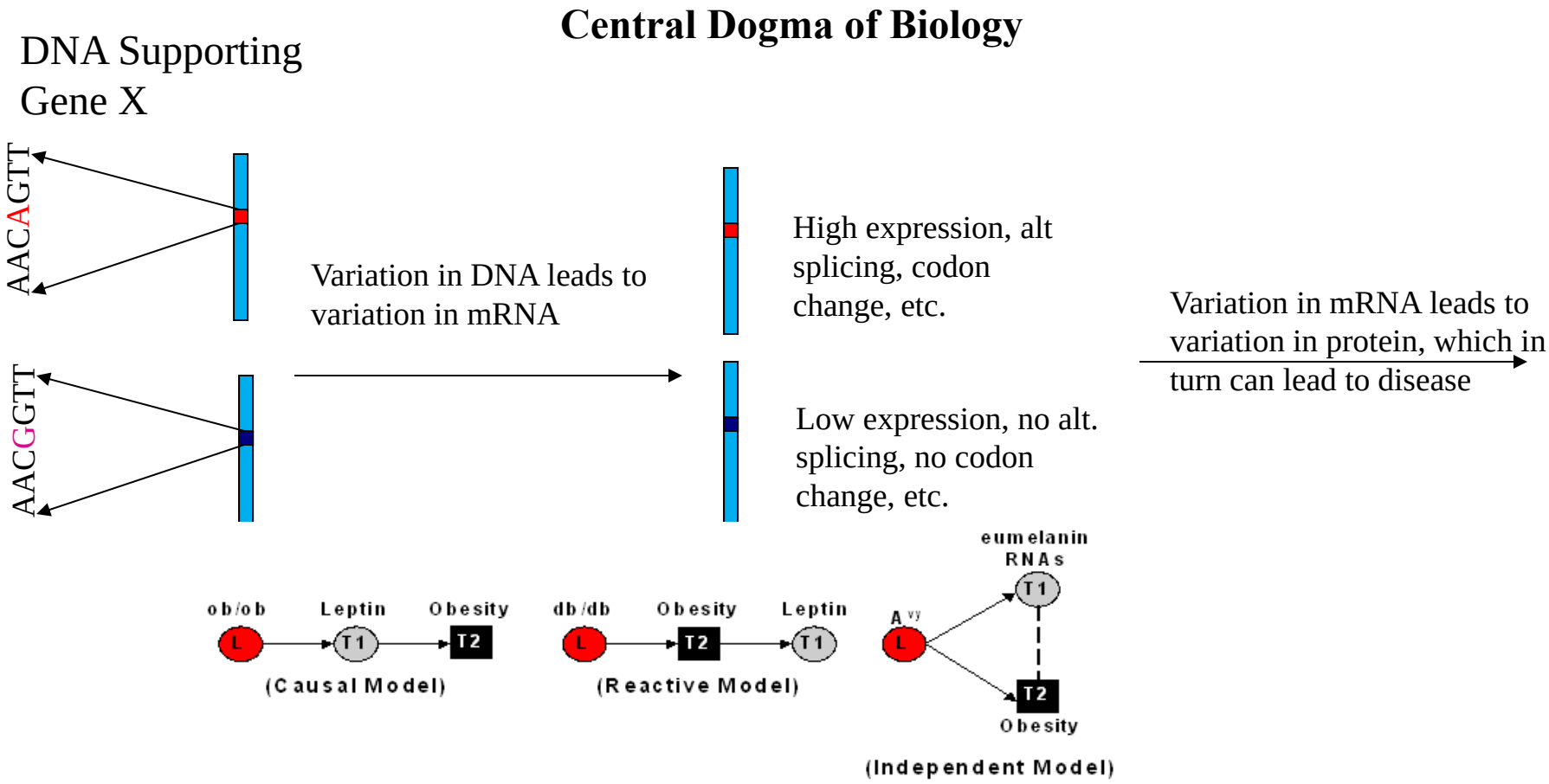




Causal inference: genetics

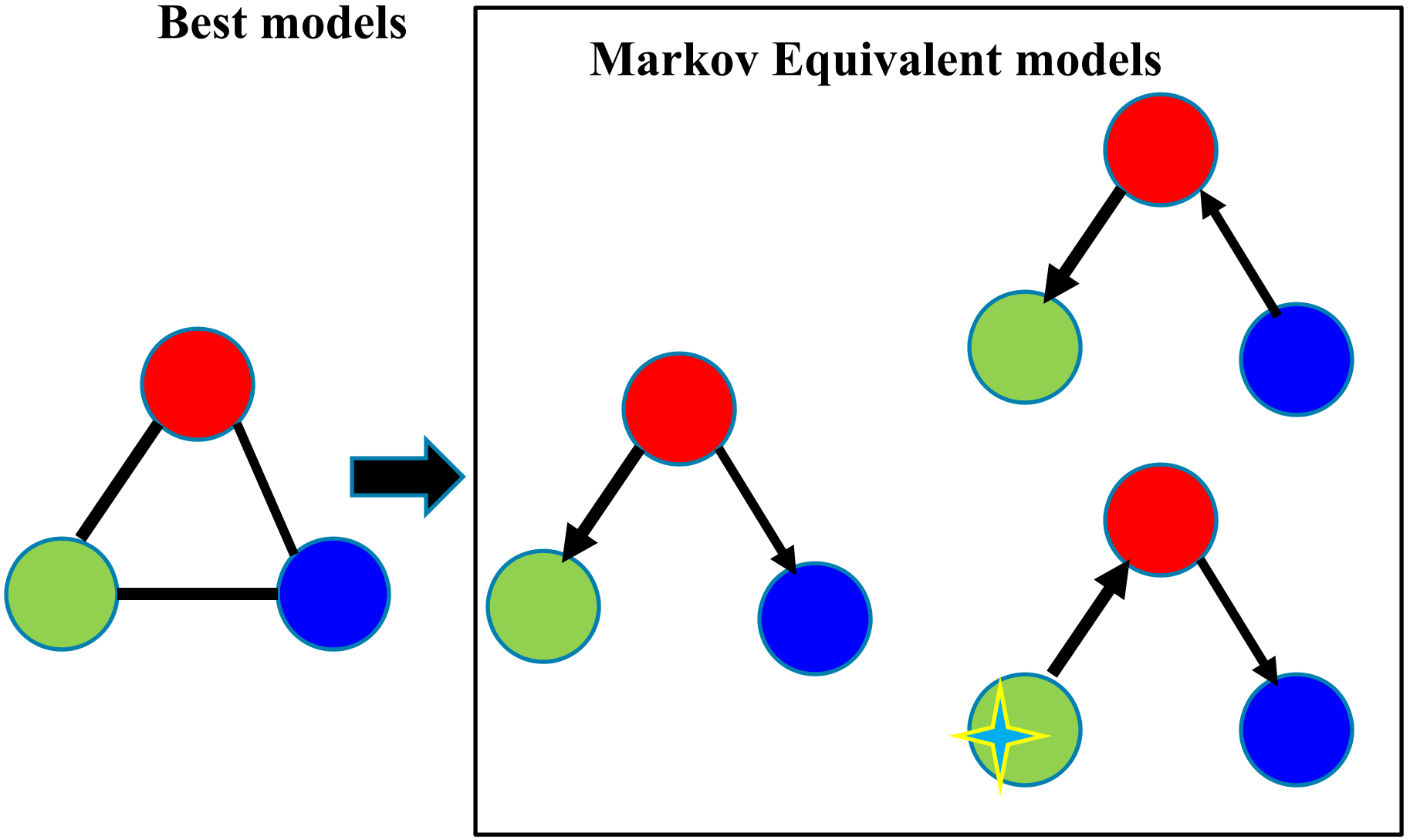
Perturbations with a causal anchor

--Natural variation in a segregating population provides the same type of causal anchor

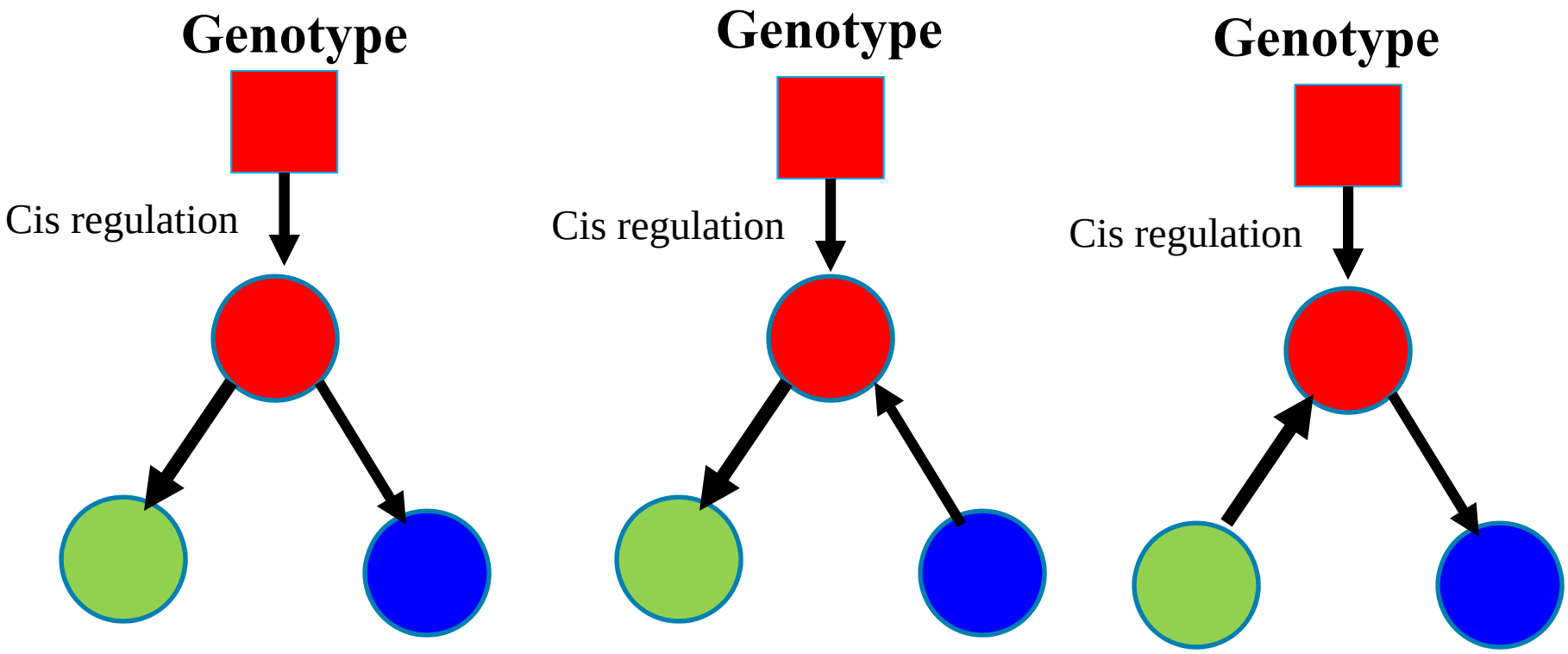


Schadt et al. Nature Genetics (2005)

A Bayesian network approach:

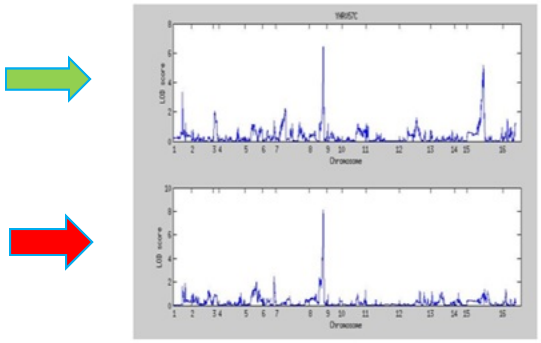


A Bayesian network approach:

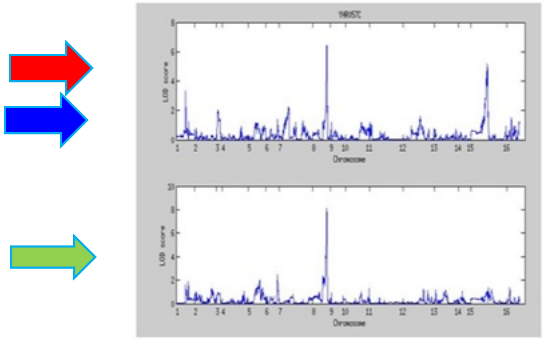


Not Markov Equivalent models

Yeast data

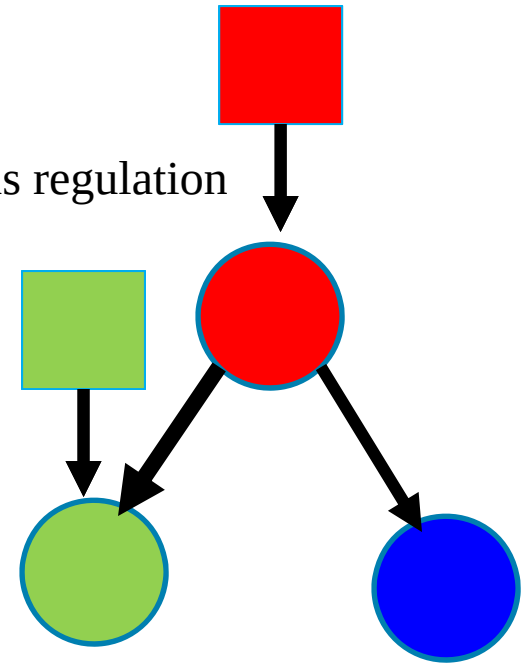


Yeast data



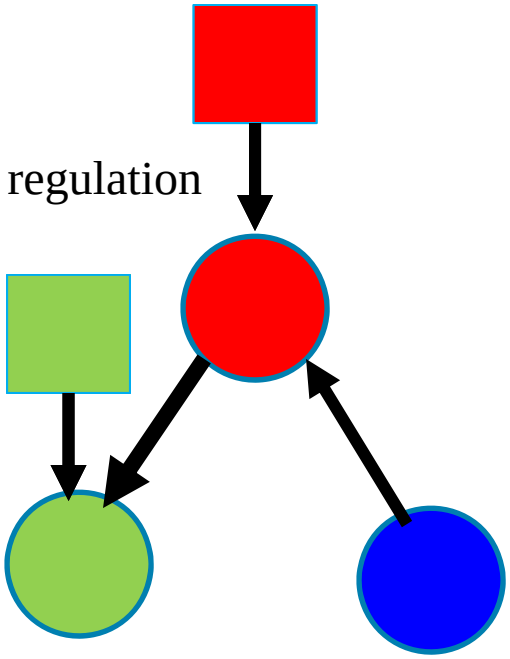
Genotype

Cis regulation



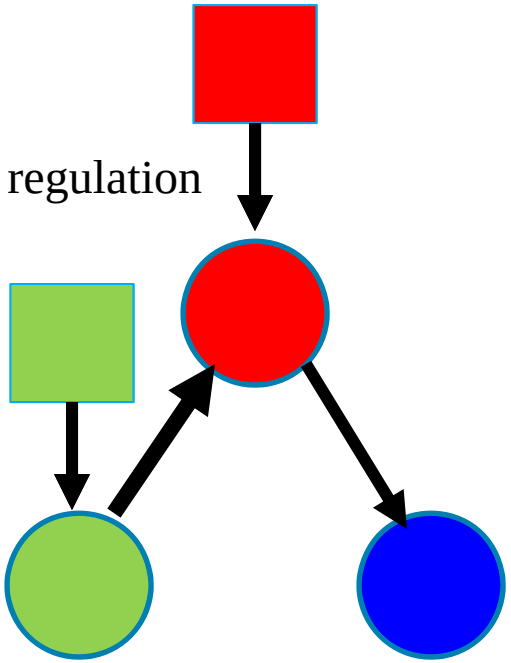
Genotype

Cis regulation



Genotype

Cis regulation



Not Markov Equivalent models

Structure priors based on causality

- Estimate confidence of causality
  - Bootstrap samples for 200 times
  - Factions of causal, reactive, independent calls

- The pair is independent

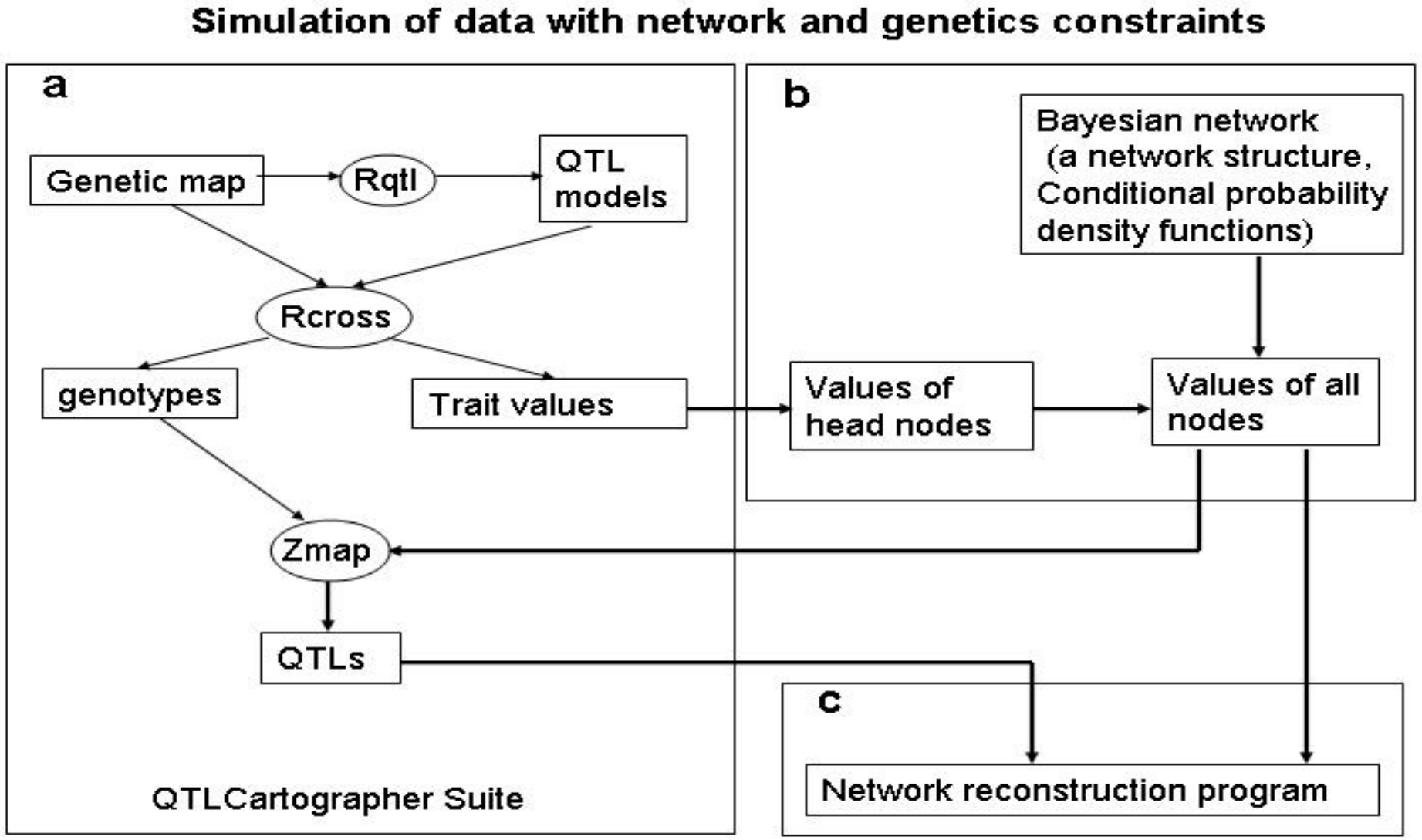
$$p(X_a -> X_b) = 1 - \frac{\sum_{i \in PE} p(X_a \perp X_b \mid l_i)}{\sum_{i \in PE} 1}$$

- The pair is causa/reactive

$$p(X_a -> X_b) = \frac{2 * \sum_i p(X_a -> X_b \mid l_i)}{\sum_i p(X_a -> X_b \mid l_i) + p(X_b -> X_a \mid l_i)}$$

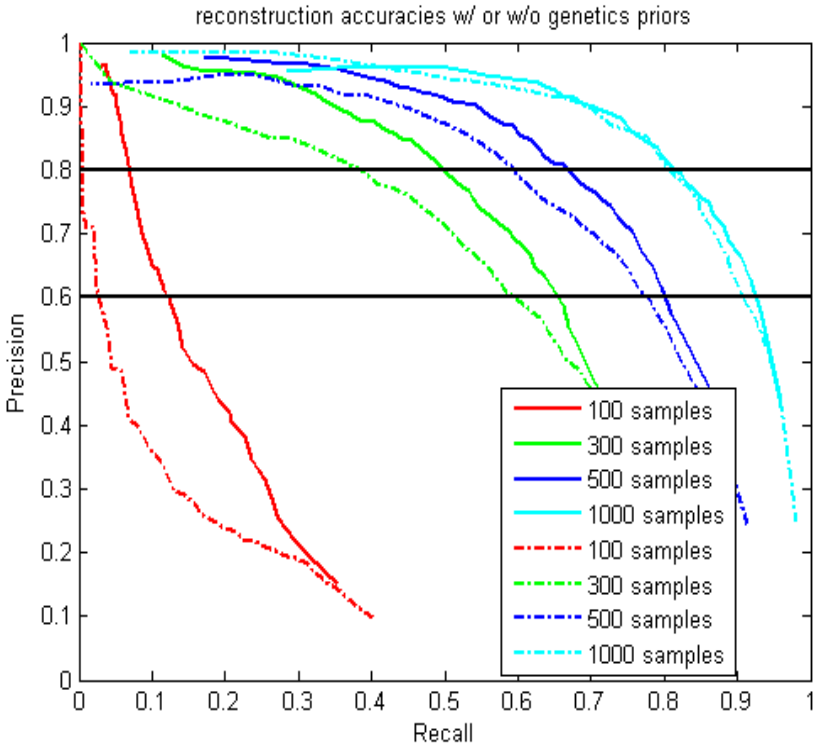
Zhu et al., PLoS CompBio, 2007

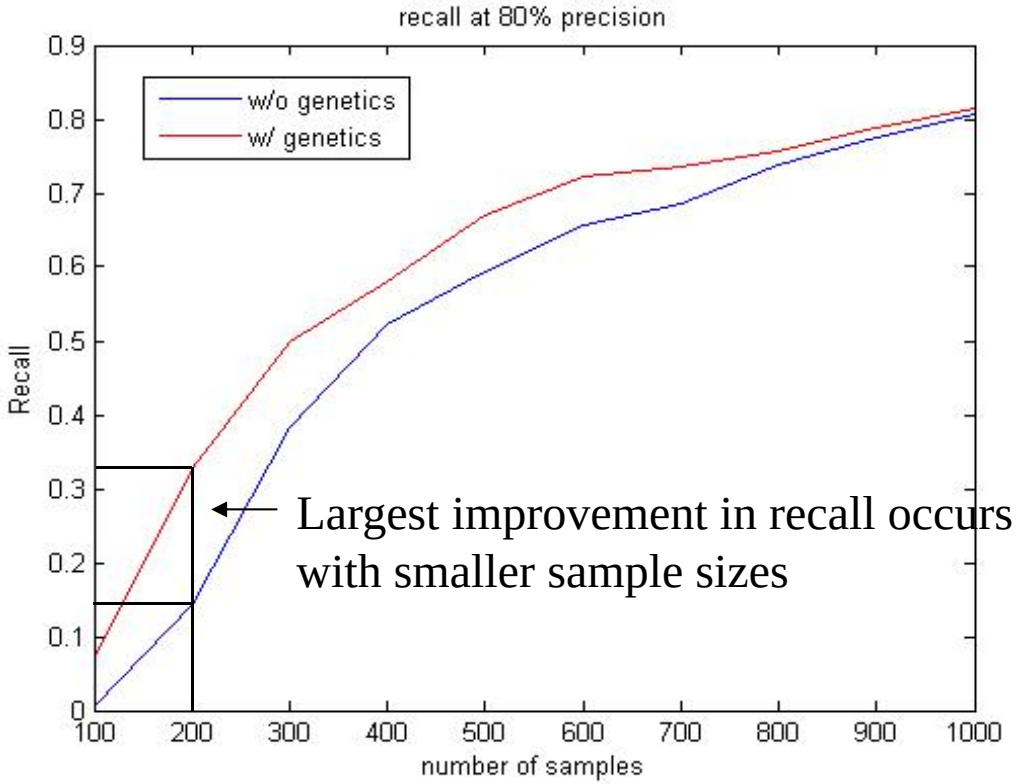
Bayesian Network: a simulation study



Zhu et al., PLoS CompBio, 2007

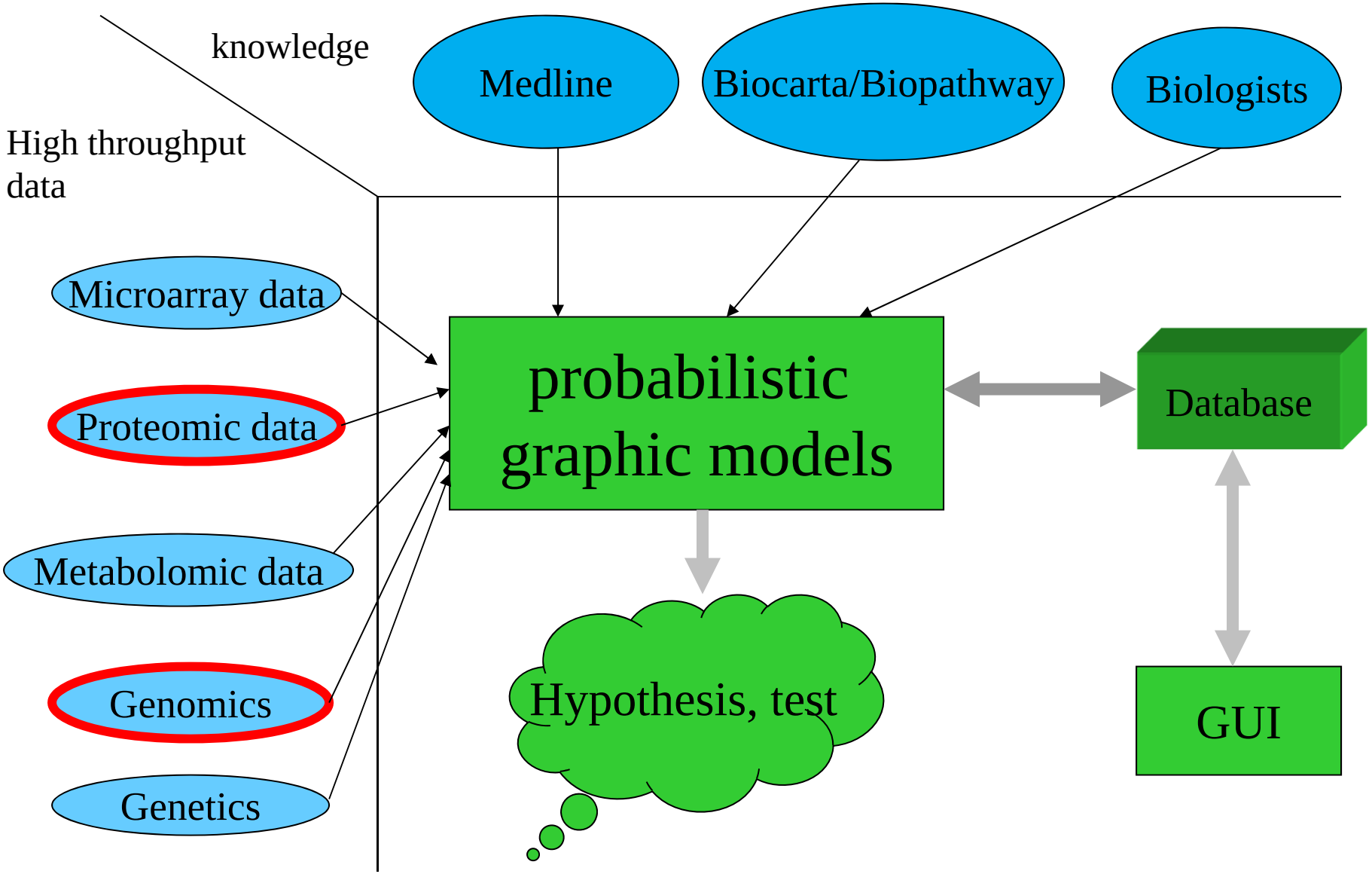
Bayesian network: Genetics information is critical when sample size is small



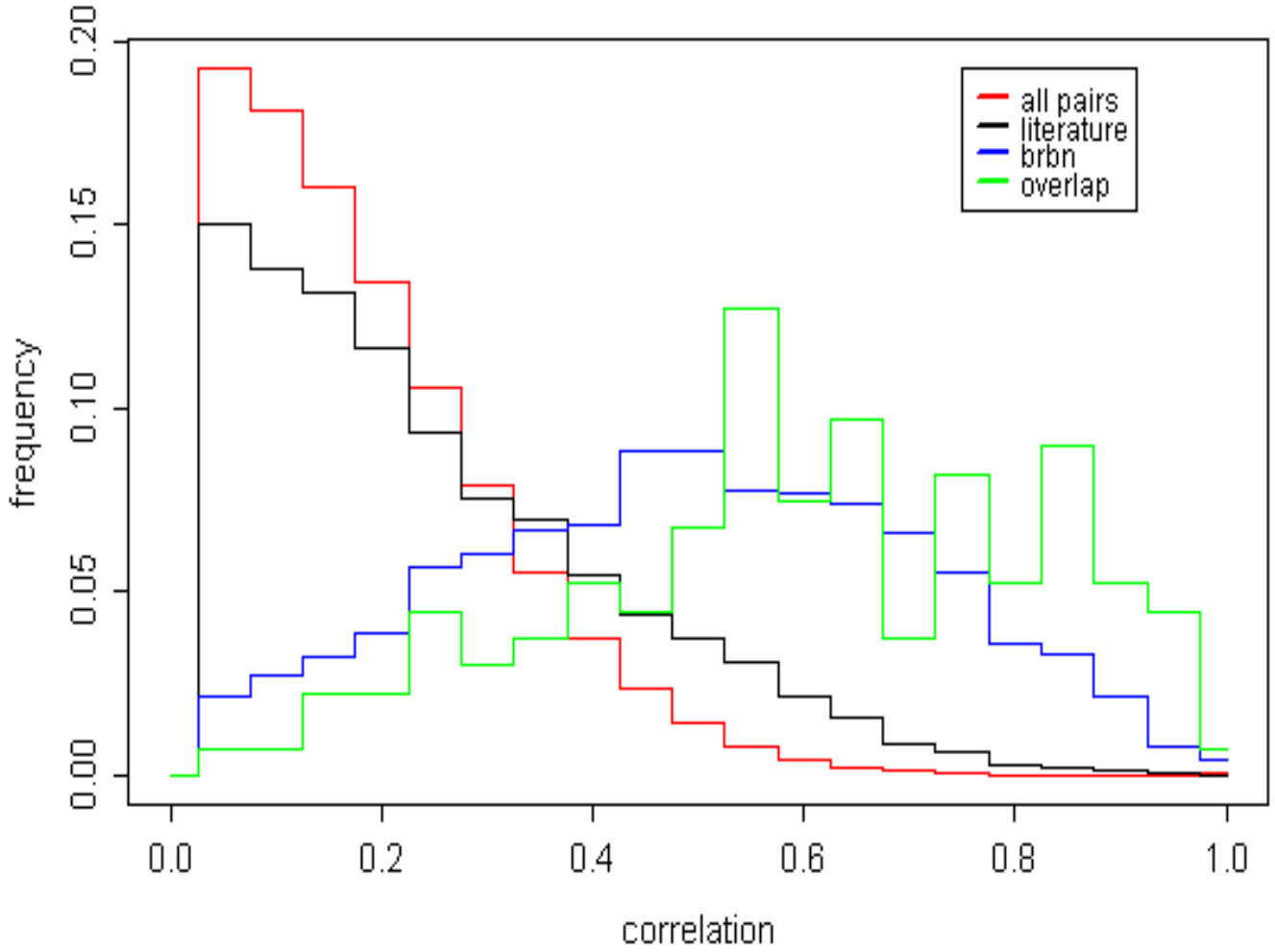


Zhu et al., PLoS CompBio, 2007

A framework for data integration

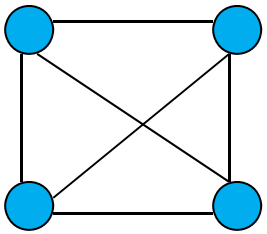


Bayesian network: PPI

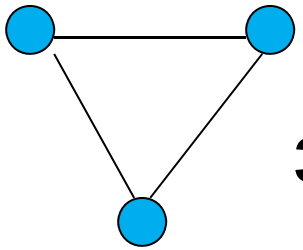


Zhu J et al, Nature Genetics, 2008

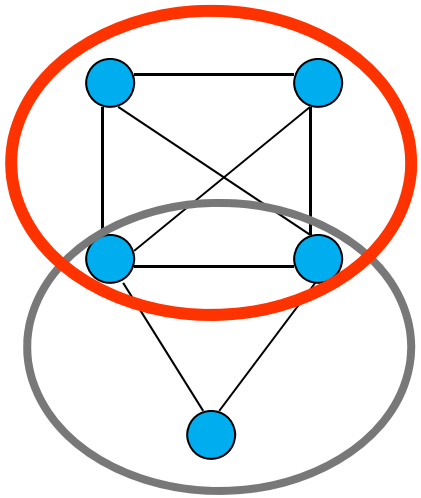
Bayesian network: PPI



4-clique



3-clique



4-clique

3-clique

Clique community
(partial clique)

Zhu J et al, Nature Genetics, 2008

Bayesian network: Transcription Factors

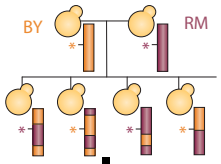
Introducing *scale-free priors* for TF or protein complex

$$p(T->g) \propto w(T)$$

$$w(T) = \log\left(\sum_{g_i \in R} \left| r(T, g_i) \right| > r_{cutoff} \right)$$

Zhu J et al, Nature Genetics, 2008

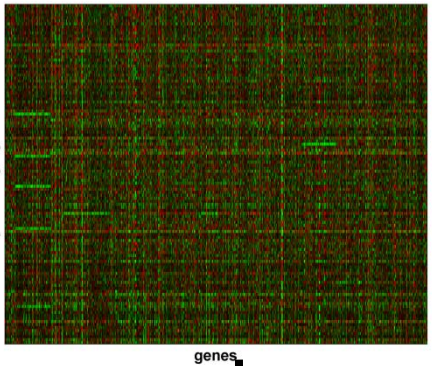
Yeast segregants



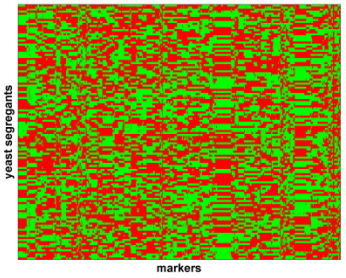
Public databases

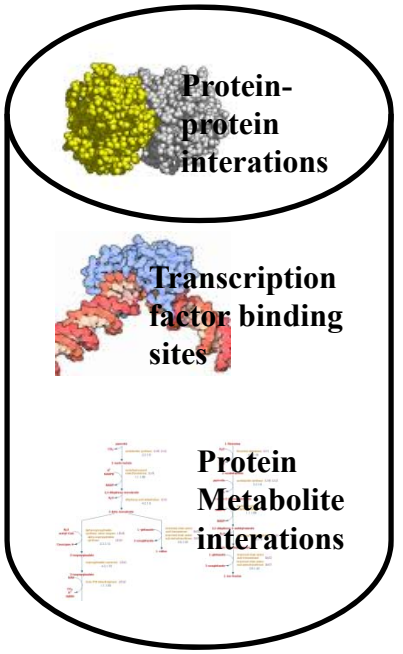
Synthetic complete medium
Logorithm growth

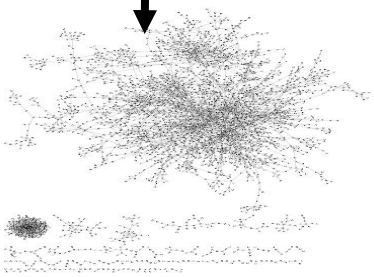
Gene expression



genotypes







Bayesian network

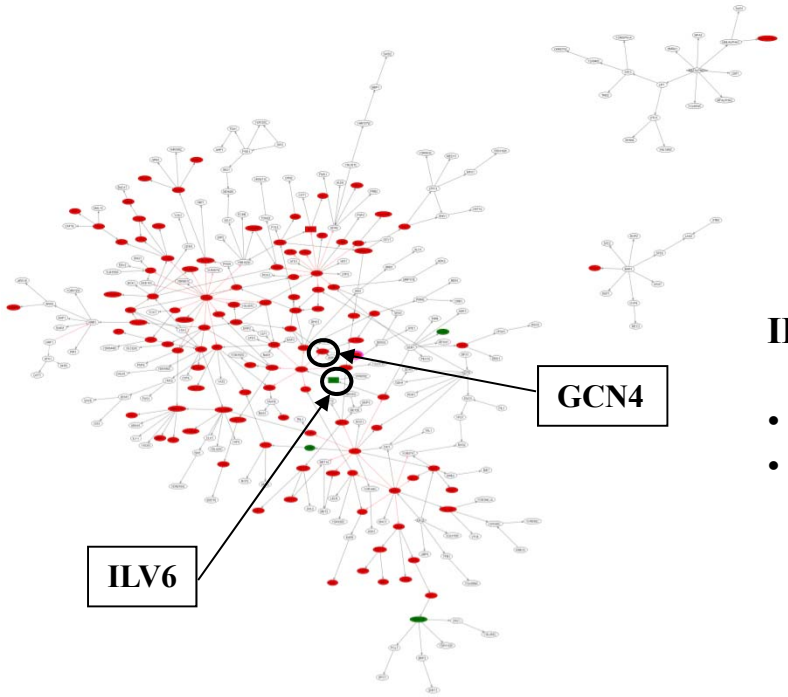
Zhu J et al, Nature Genetics, 2008

Integration improves network qualities

BN	KO data	GO terms	TF data
w/o any priors	125	55	26
w/ genetics priors	139	59	34
w/ genetics, TF and PPI priors	152	66	52

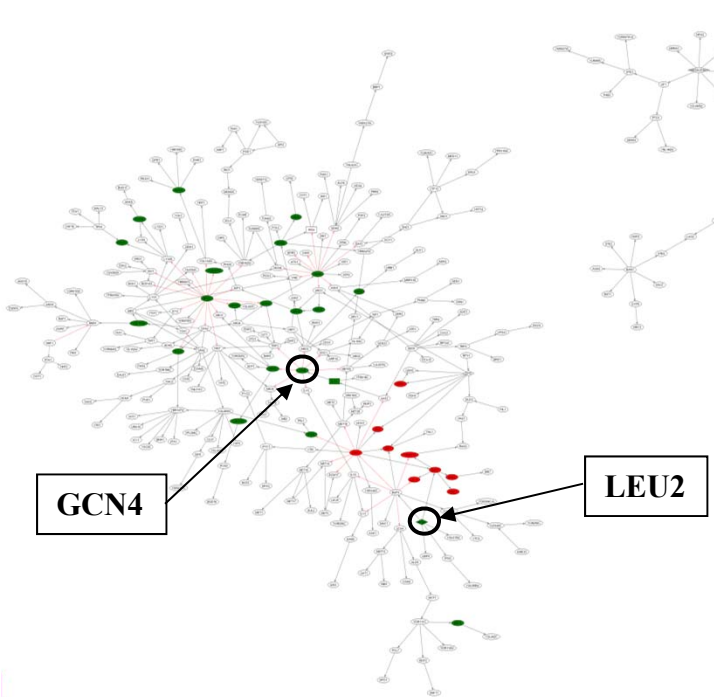
Zhu J et al, Nature Genetics, 2008

Prospective validation is the gold standard



ILV6 gives rise to large expression signature

- ILV6 KO sig enriched (p~10E-52)
- GCN4 upregulated in ILV6 KO → large signature



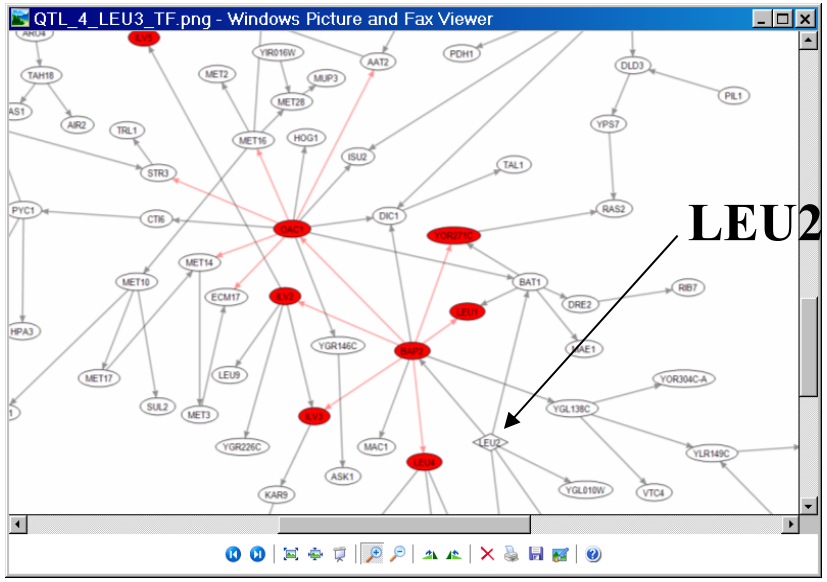
LEU2 KO gives rise to small expression signature

- LEU2 KO sig enriched (p~10E-18)
- GCN4 downregulated in LEU2 KO → small signature

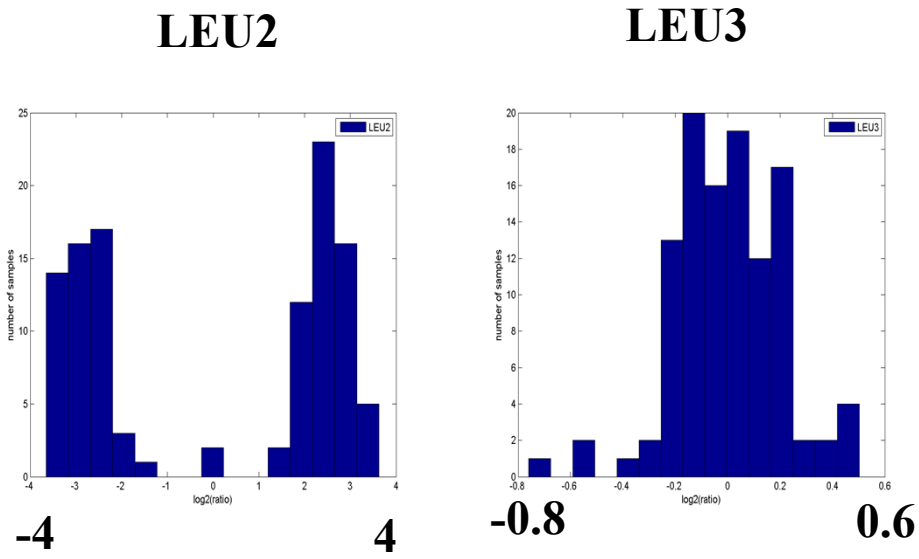
Zhu J et al, Nature Genetics, 2008

How does LEU2 affect LEU3 activity?

LEU3 binding sites

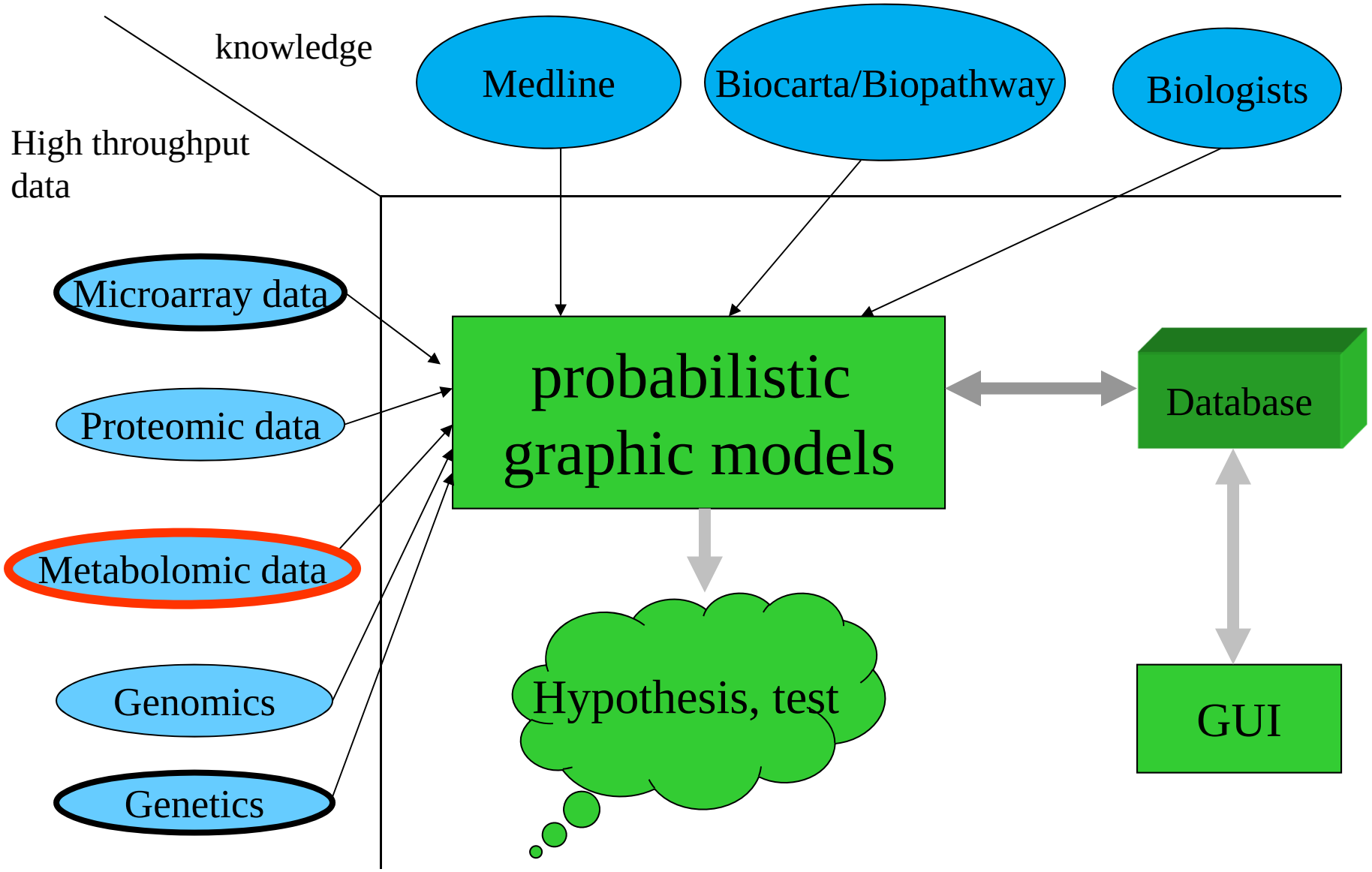


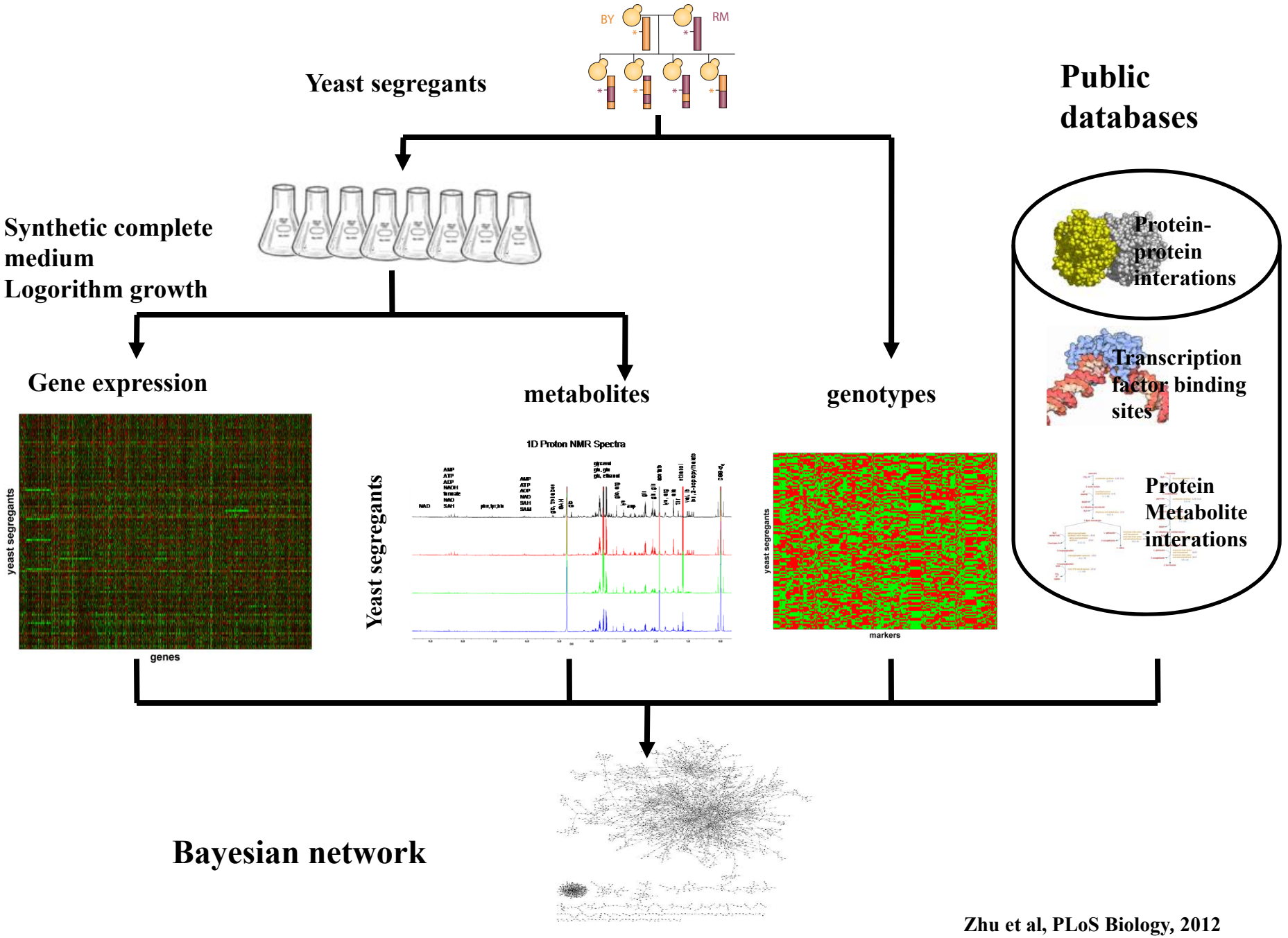
mRNA expression



Surrogate marker for Leu3p activity

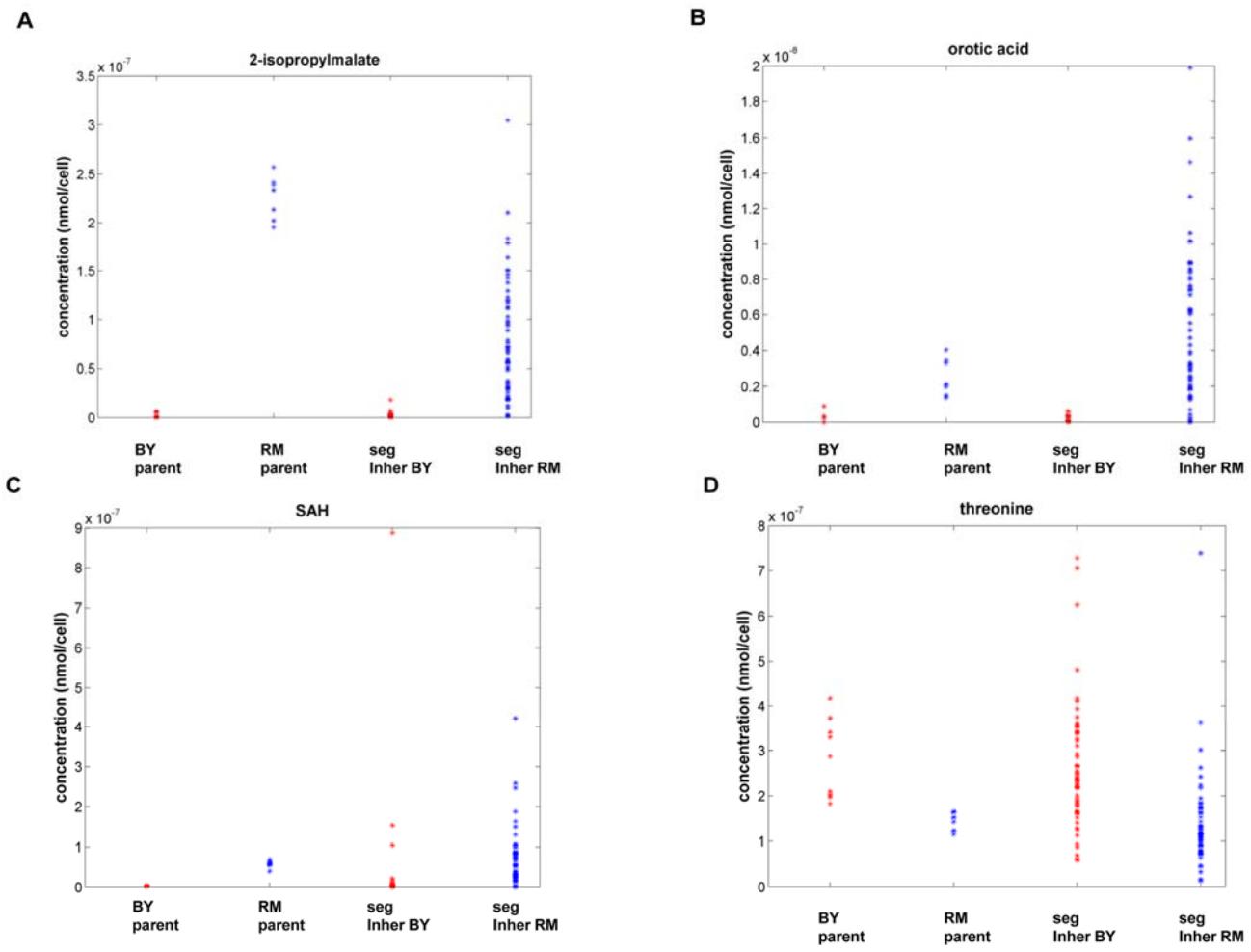
A framework for building causal networks





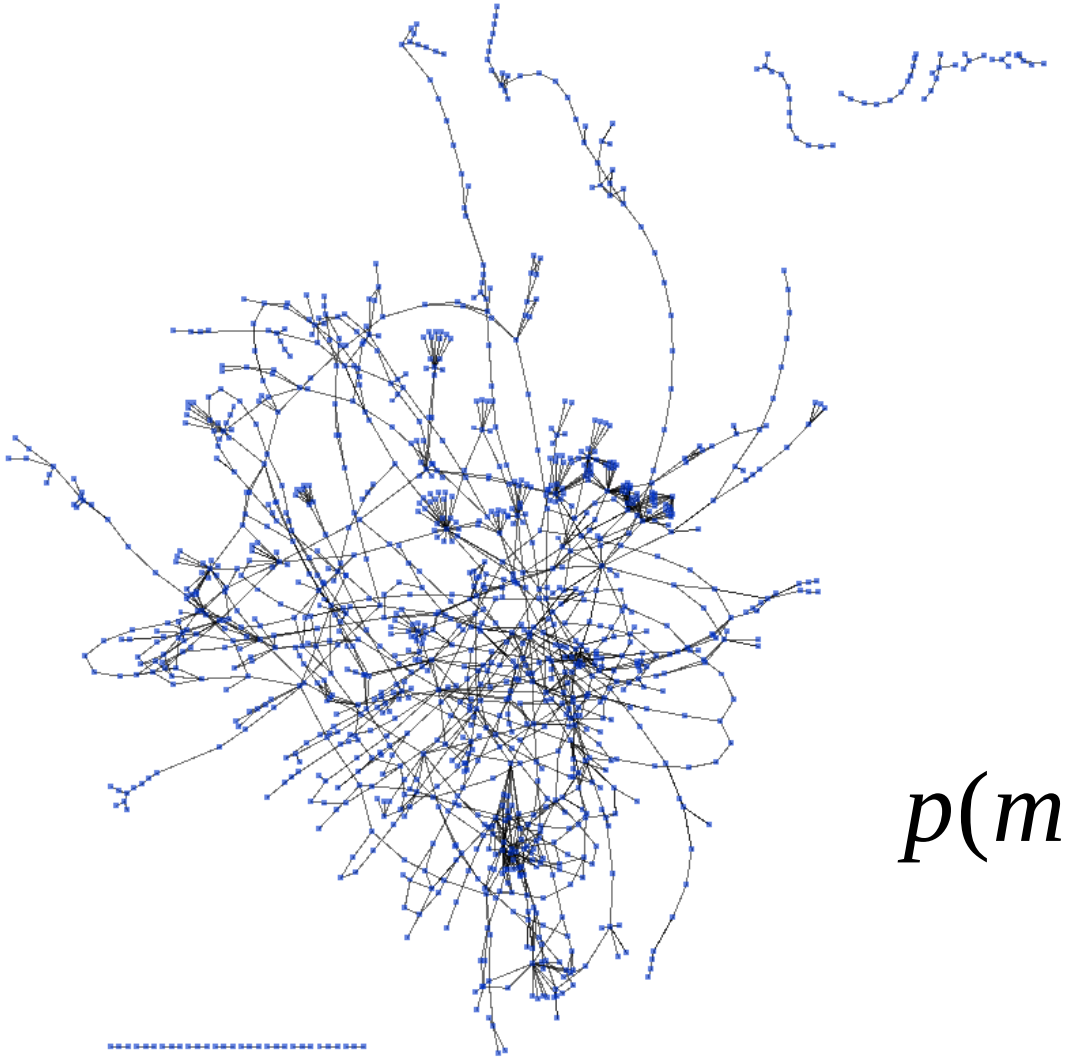
Zhu et al, PLoS Biology, 2012

Metabolite abundance is under genetic control



Zhu et al, PLoS Biology, 2012

KEGG biochemical pathways

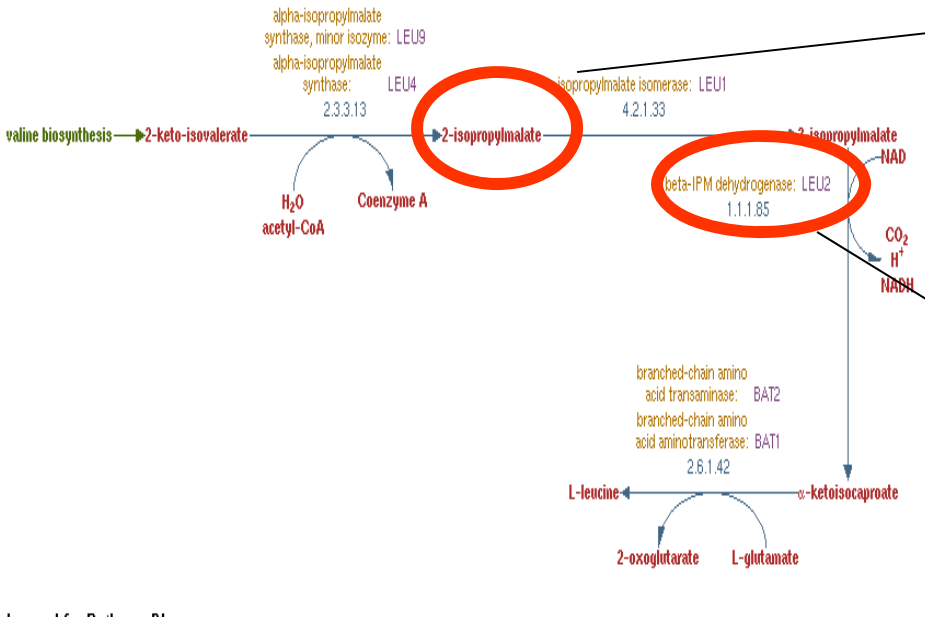


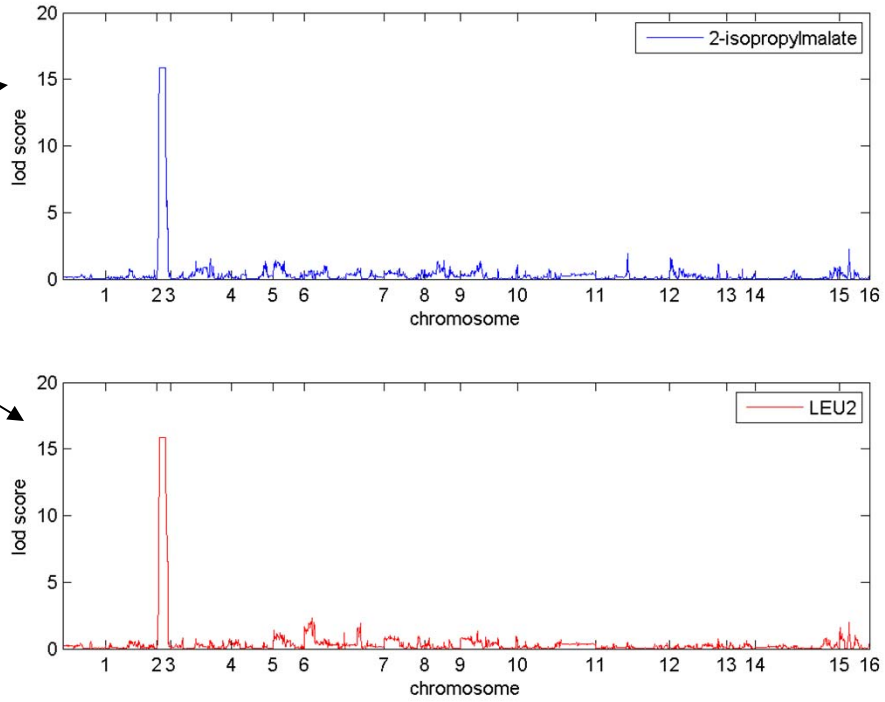
$$p(m->e) \propto e^{-\lambda d_{m,e}}$$

Zhu et al, PLoS Biology, 2012

LEU2 mRNA is causal to 2-isopropylmalate

KEGG pathway

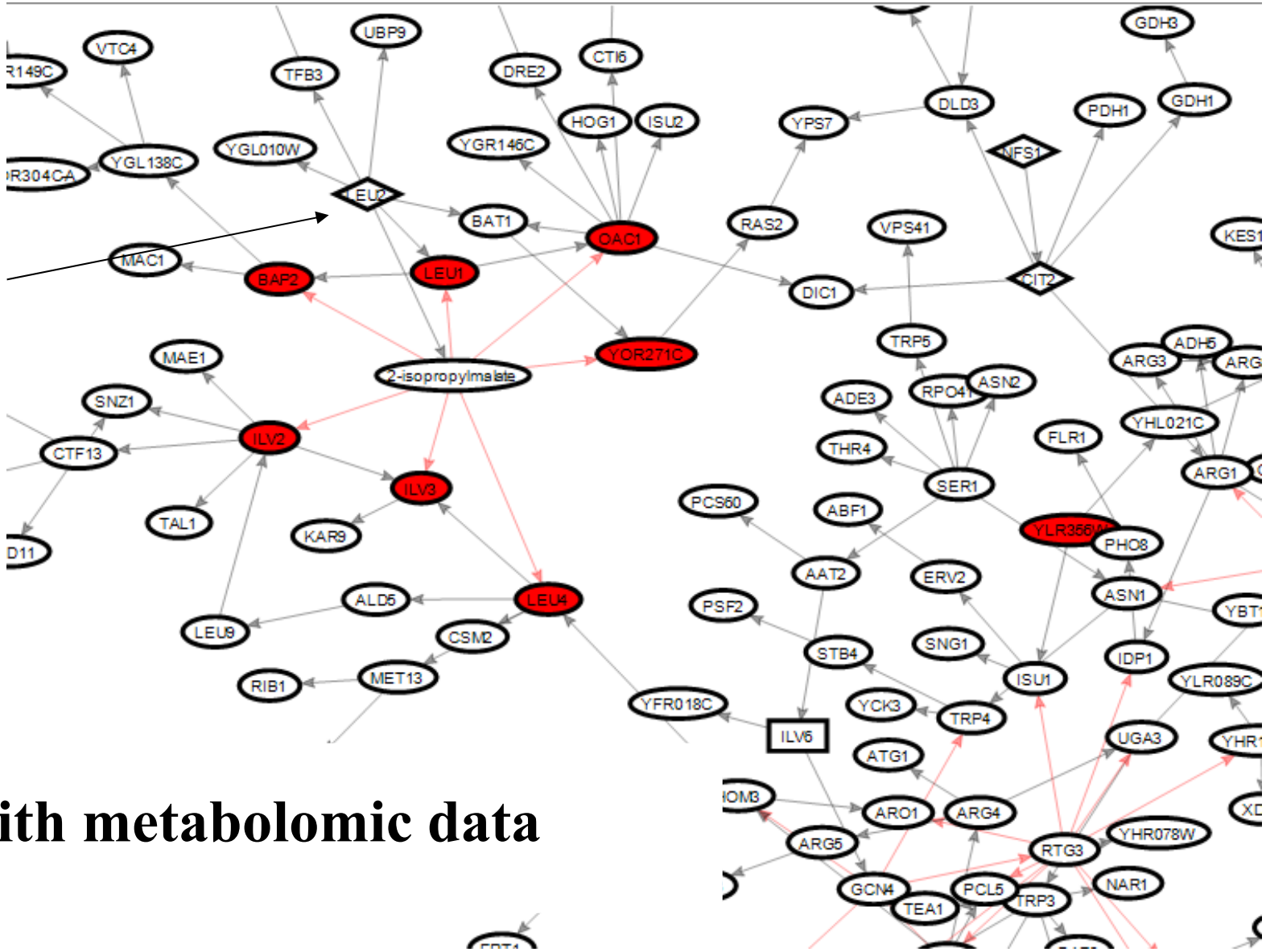




Zhu et al, PLoS Biology, 2012

LEU3 binding site

LEU2



With metabolomic data

Zhu et al, PLoS Biology, 2012

LEU3 regulation

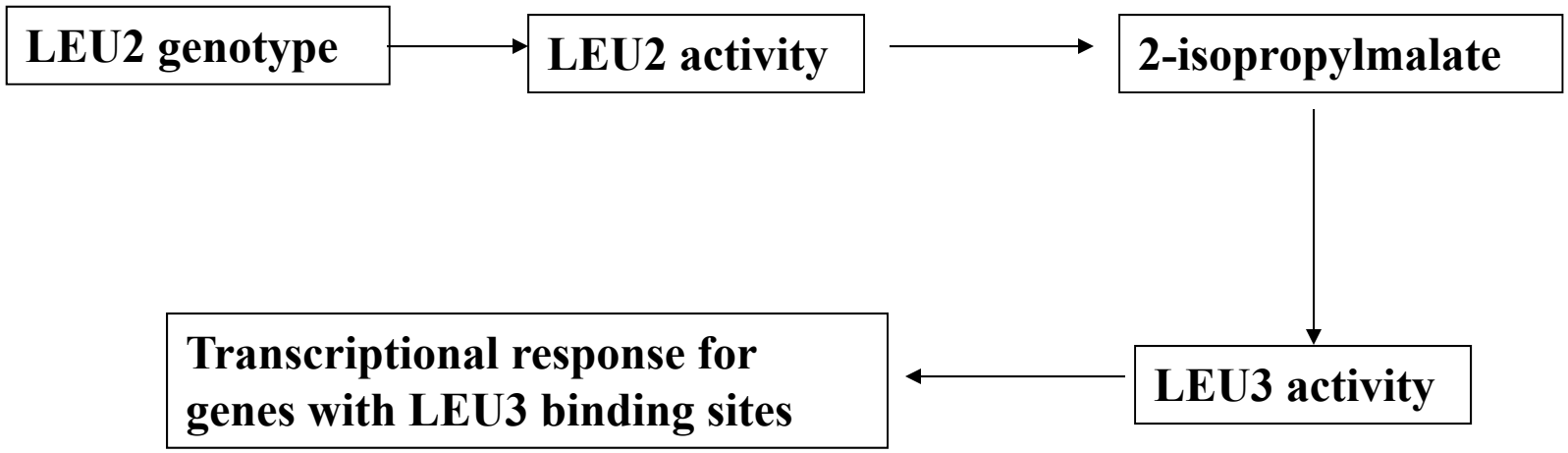
- **The activity of *Leu3*p is positively regulated by alpha-isopropylmalate (IPM), the product of the first step in leucine biosynthesis**

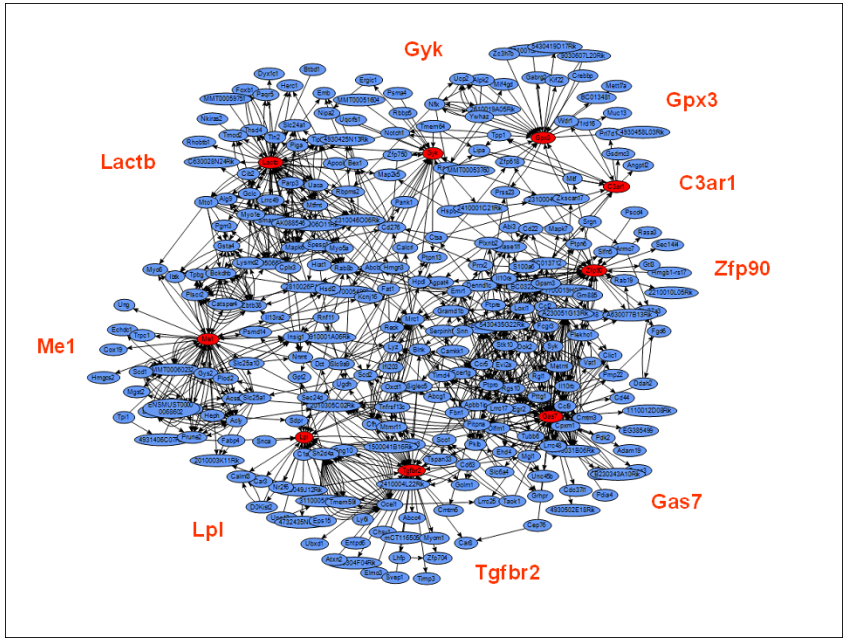
  Sze JY, *et al.* (1992) In vitro transcriptional activation by a metabolic intermediate: activation by Leu3 depends on alpha-isopropylmalate. *Science* 258(5085):1143-5

- **The degree of activation by *Leu3*p is *Leu3*p concentration dependent, and it has been shown that *LEU3* gene expression is regulated by general amino acid control, which is mediated by the GCN4 transcription factor**

  Zhou K, *et al.* (1987) Structure of yeast regulatory gene LEU3 and evidence that LEU3 itself is under general amino acid control. *Nucleic Acids Res* 15(13):5261-73

2-isopropylmalate: mechanism of causal regulator LEU2





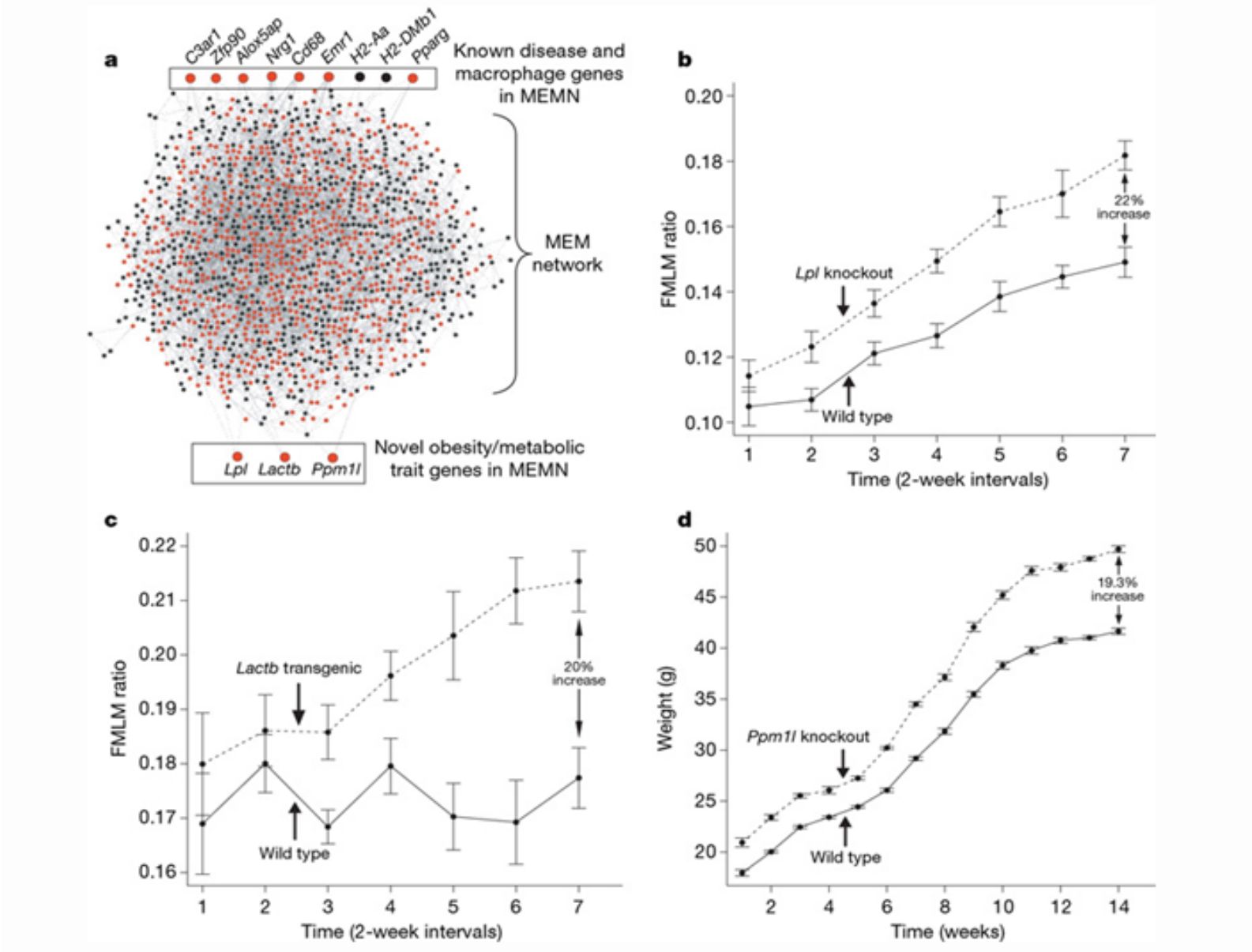
Networks facilitate direct identification of genes that are causal for disease

Yang et al, Nature Genetics (2009)

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

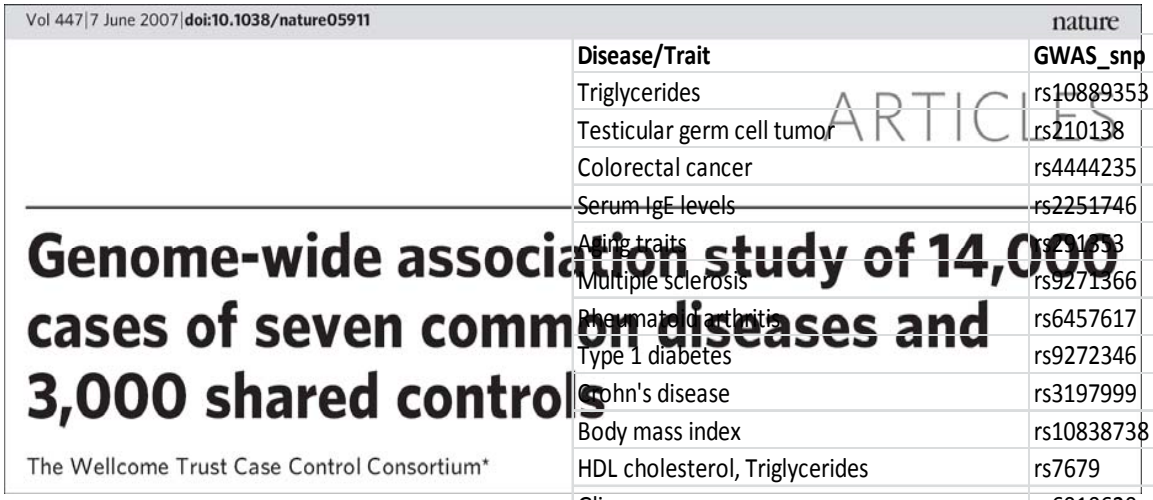
Schadt et al. Nature Genetics (2005)

Multiple genes in a network causing diseases!



Emilsson et al, Nature 2008; Chen Y et al. Nature, 2008

Guilt by association

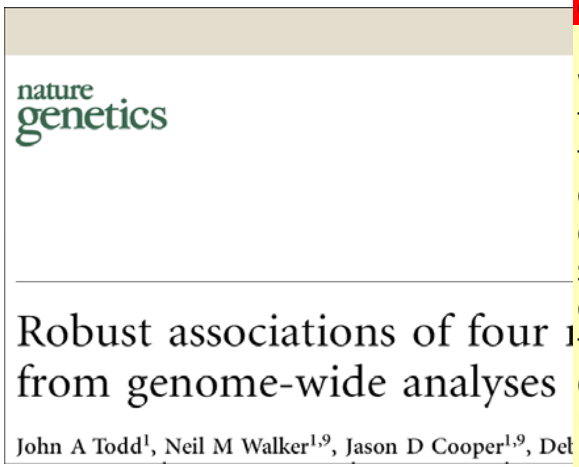

Vol 447|7 June 2007|doi:10.1038/nature05911

nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

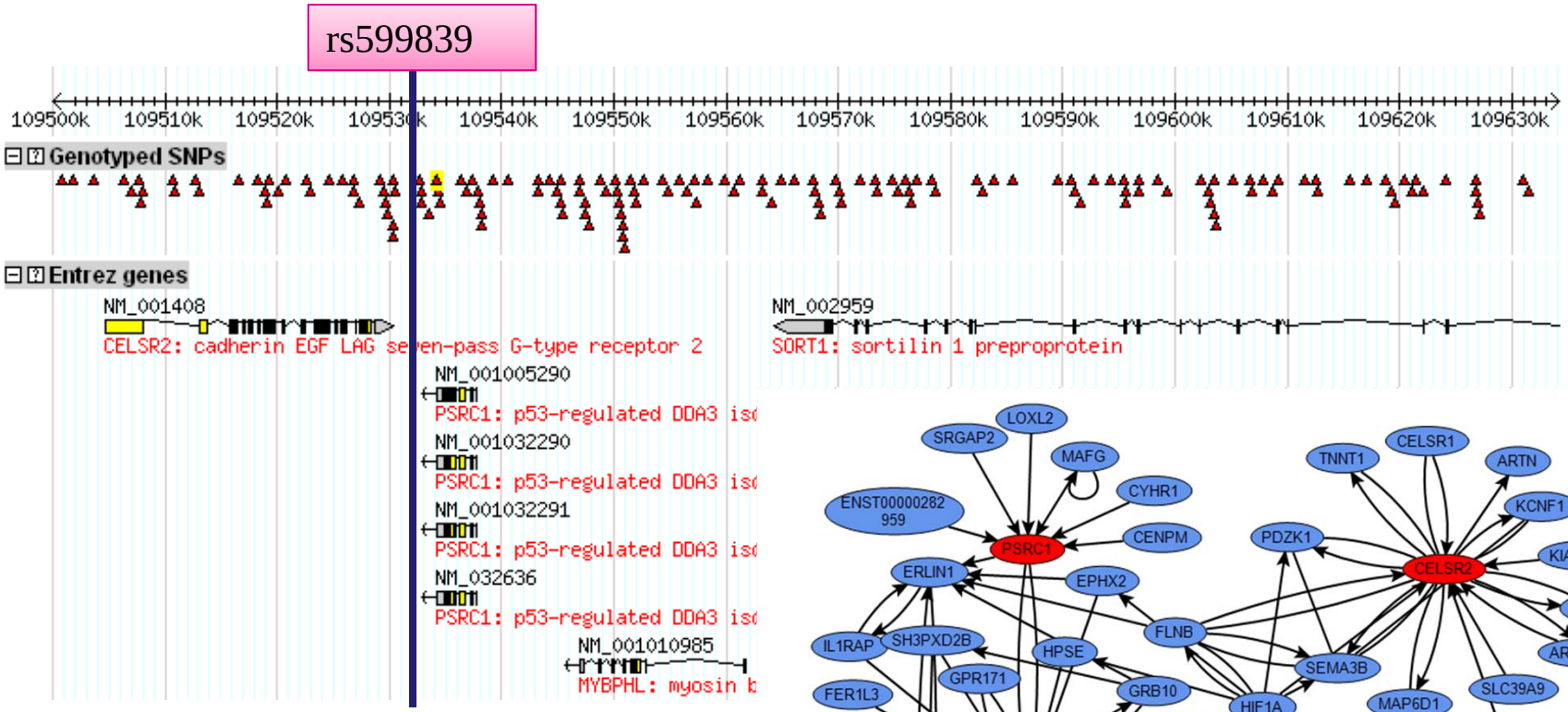

nature genetics

Robust associations of four
from genome-wide analyses

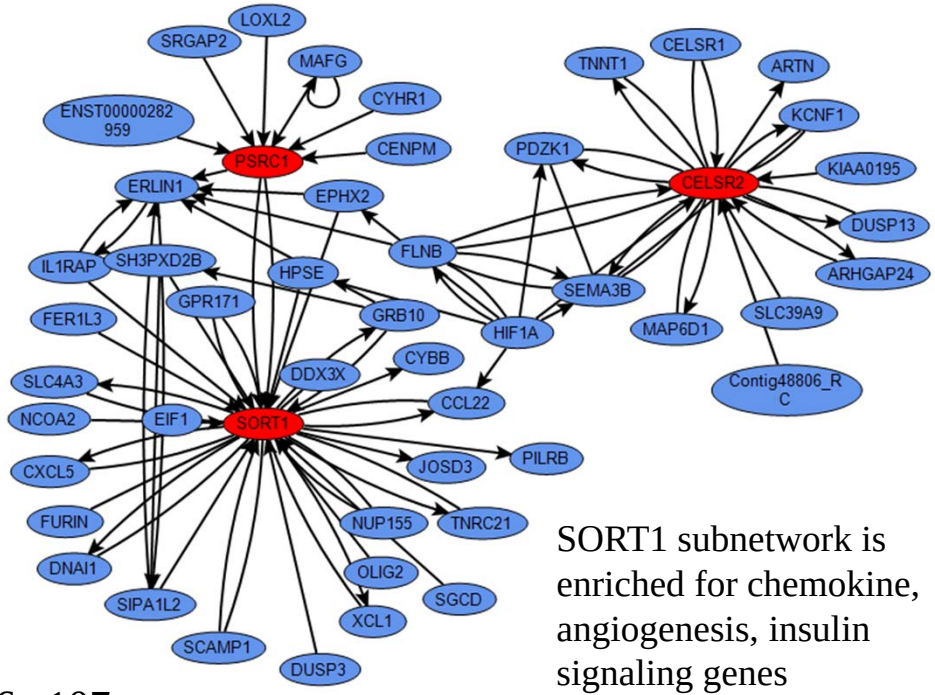
John A Todd[1], Neil M Walker[1,9], Jason D Cooper[1,9], Deb

Disease/Trait	GWAS_snp	Reported Gene(s)	eSNP_gene
Triglycerides	rs10889353	ANGPTL3	ANGPTL3
Testicular germ cell tumor	rs210138	BAK1	BAK1
Colorectal cancer	rs4444235	BMP4	BMP4
Serum IgE levels	rs2251746	FCER1A	FCER1A
Aging traits	rs2216573	GNG4	GNG4
Multiple sclerosis	rs9271366	HLA-DRB1	HLA-DRB1
Rheumatoid arthritis	rs6457617	MHC	HLA-DQA1
Type 1 diabetes	rs9272346	MHC	HLA-DQB1
Crohn's disease	rs3197999	MST1	MST1
Body mass index	rs10838738	MTCH2	MTCH2
HDL cholesterol, Triglycerides	rs7679	PLTP	PLTP
Glioma	rs6010620	RTEL1	RTEL1
Attention deficit hyperactivity disorder	rs260461	ZNF544	ZNF544
Coronary disease, LDL	rs599839	CELSR2, PSRC1, MYBPHL	SORT1, PSRC1, CELSR2
Lung cancer	rs8034191	CHRNA3, CHRNA5, PSMA4, LOC123688	PSMA4
Weight	rs2844479	AIF1, NCR3	BAT3
Type 1 diabetes	rs7804356	Intergenic	SKAP2
Testicular germ cell tumor	rs4699052	Intergenic	LOC56898
Crohn's disease	rs6596075	Intergenic	SLC22A5
Cognitive test performance	rs2832077	Intergenic	CCT8
Systemic lupus erythematosus	rs10798269	Intergenic	C1orf9
Colorectal cancer	rs4779584	Intergenic	GREM1
Type 1 diabetes	rs1701704	RAB5B, SUOX, IKZF4, ERBB3, CDK2	RPS26
QT interval	rs4725982	KCNH2	IAN4L1
Intracranial aneurysm	rs700651	BOLL, PLCL1	PLCL1
Body mass index	rs7498665	SH2B1	EIF3C
Rheumatoid arthritis	rs881375	TRAF1, C5	TRAF1
Attention deficit hyperactivity disorder	rs2842643	TFEB	UNC5CL

Our Data Supports SORT1 as the Strongest Candidate Gene



eSNP	Genes associated	p value
rs599839	SORT1	1.52E-56
rs599839	PSRC1	2.17E-53
rs599839	CELSR2	4.31E-23



SORT1 subnetwork is enriched for chemokine, angiogenesis, insulin signaling genes

Schadt et al., PLoS Biol., 2008; 6:e107

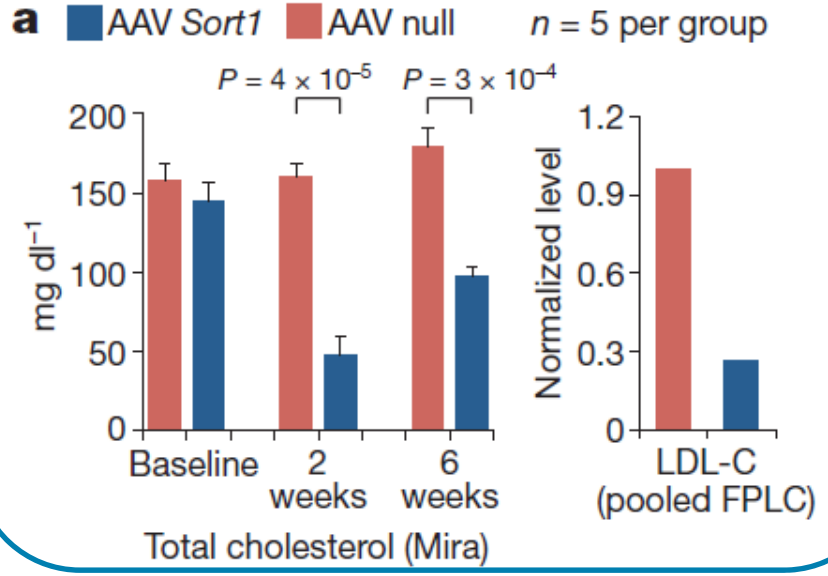
nature Vol 466|5 August 2010|doi:10.1038/nature09266

ARTICLES

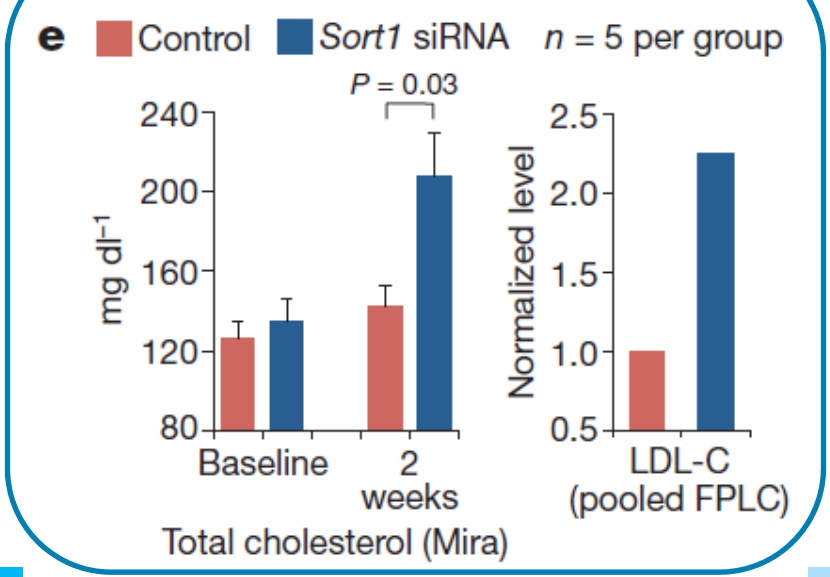
From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus

Kiran Musunuru[1,2,3]*, Alanna Strong[4]*, Maria Frank-Kamenetsky[5], Noemi E. Lee[1], Tim Ahfeldt[1,6], Katherine V. Sachs[4], Xiaoyu Li[4], Hui Li[4], Nicolas Kuperwasser[1], Vera M. Ruda[1], James P. Pirruccello[1,2], Brian Muchmore[7], Ludmila Prokunina-Olsson[7], Jennifer L. Hall[2,8], Eric E. Schadt[9], Carlos R. Morales[10], Sissel Lund-Katz[11], Michael C. Phillips[11], Jamie Wong[5], William Cantley[5], Timothy Racie[5], Kenechi G. Ejebe[1,2], Marju Orho-Melander[12], Olle Melander[12], Victor Koteliansky[5], Kevin Fitzgerald[5], Ronald M. Krauss[13], Chad A. Cowan[1,2], Sekar Kathiresan[1,2]* & Daniel J. Rader[4]*

Overexpression

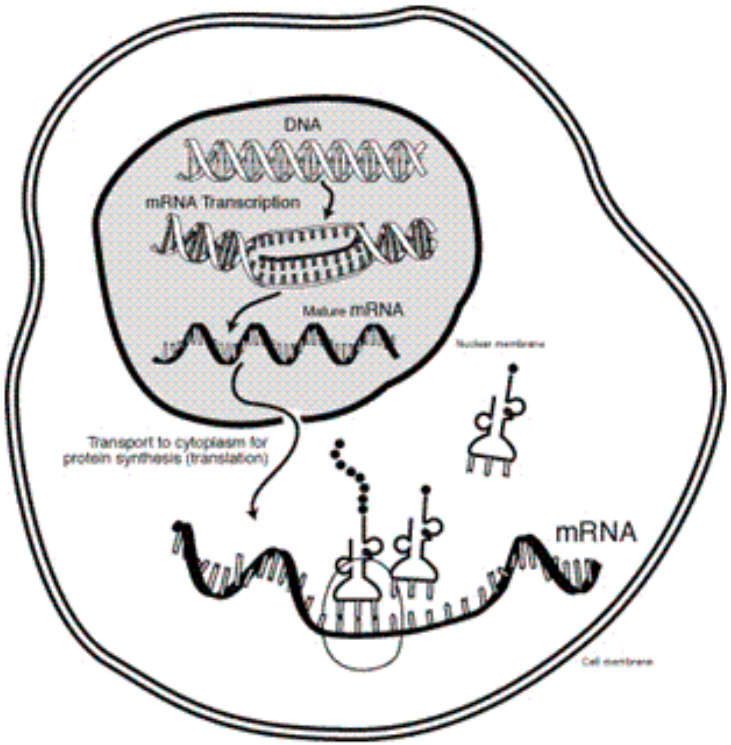


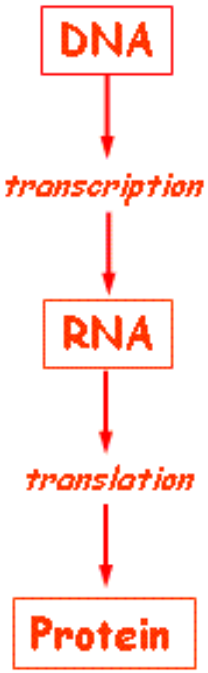
Knockdown



Why it is so hard to model biological systems?

- The more we learn, the more complicated it becomes!





It is not one gene to one protein anymore!

Epigenetic regulation : heritable changes in gene function that cannot be explained
by changes in DNA sequence

- DNA methylation
- Chromotin structure

Junk DNA?

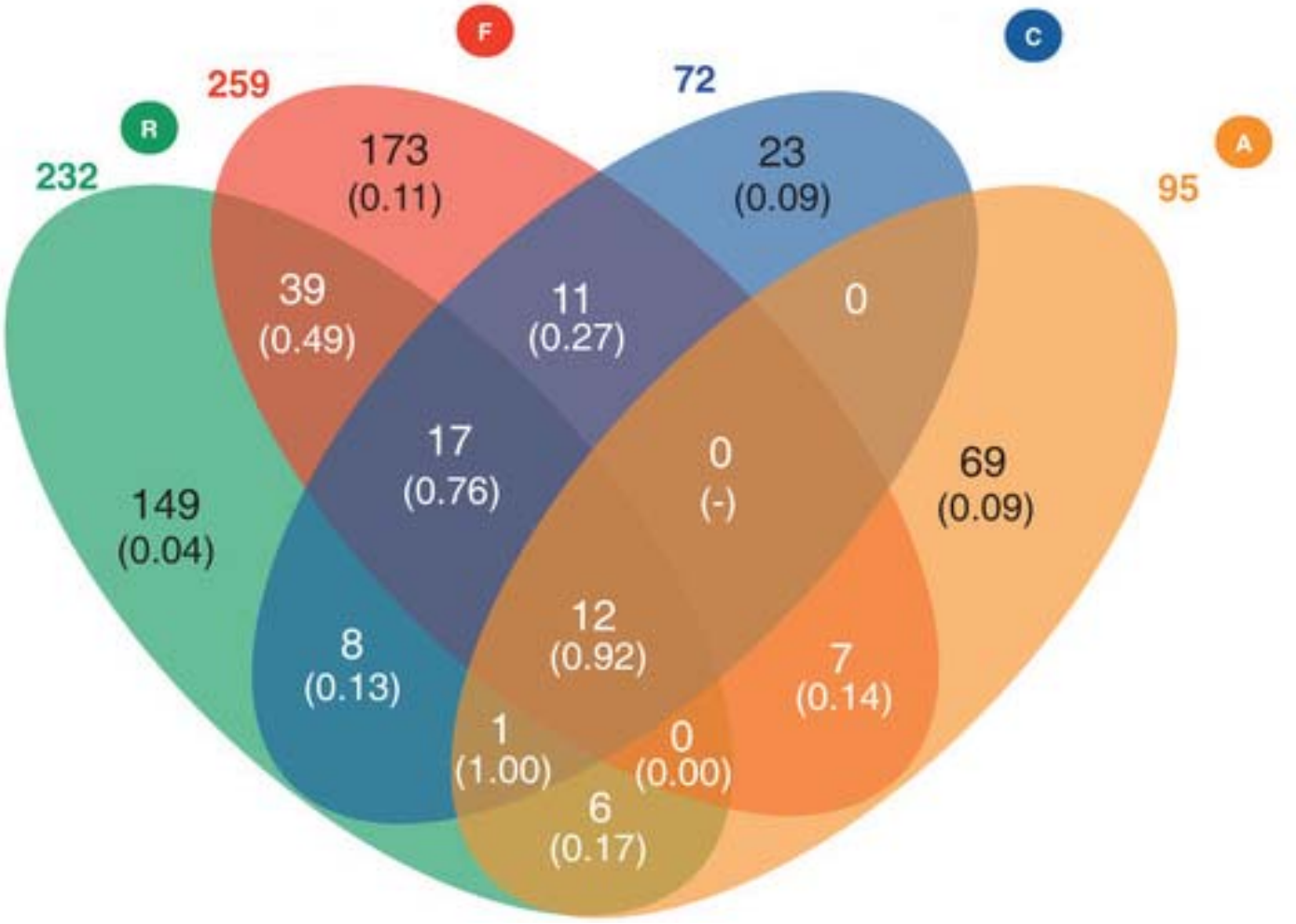
Post transcriptional regulation

- Splicing (1981)
- RNA editing (1986)
- miRNA mediated regulation (1993)

Post translational regulation

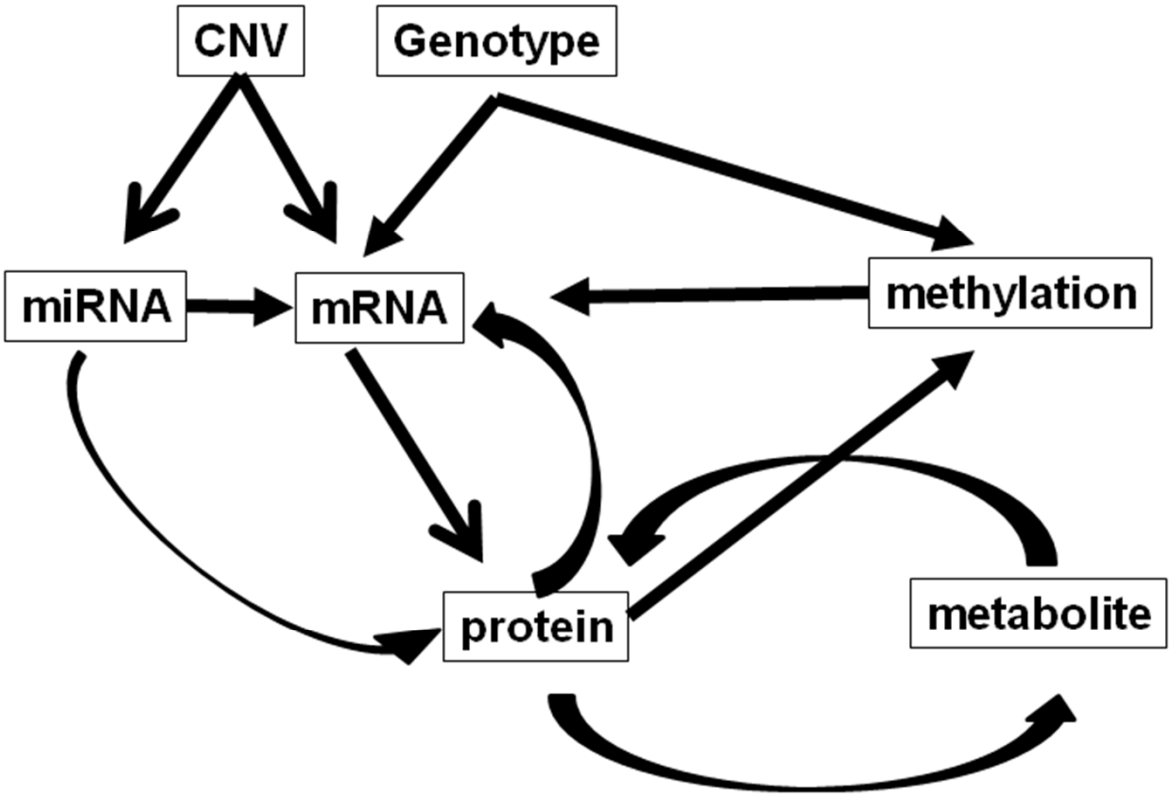
- Phosphorylation
- Glycosaltion
- acetylation

Integrating omics data



Complex diseases: observations to models

perturbations



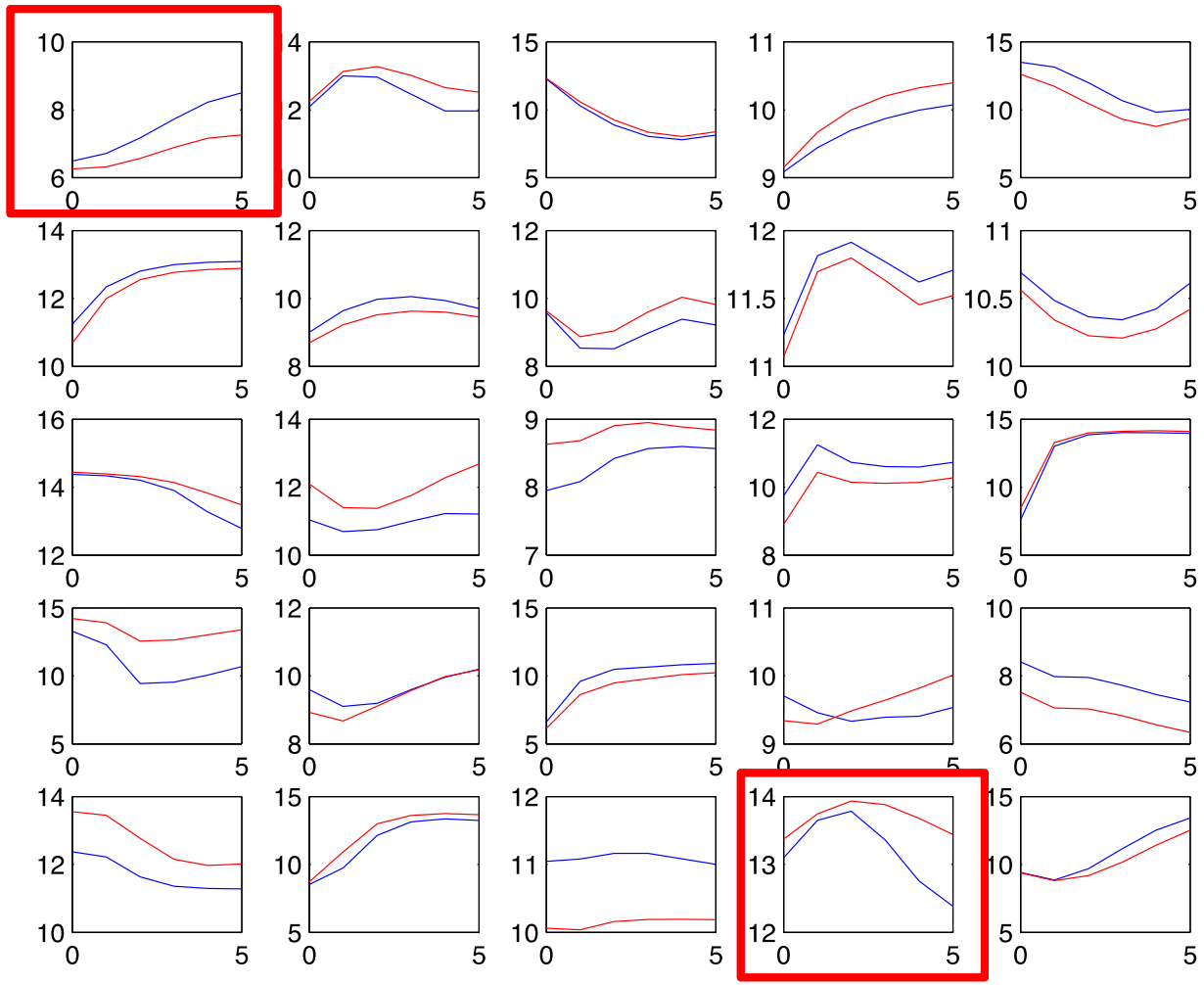
perturbations

diseases

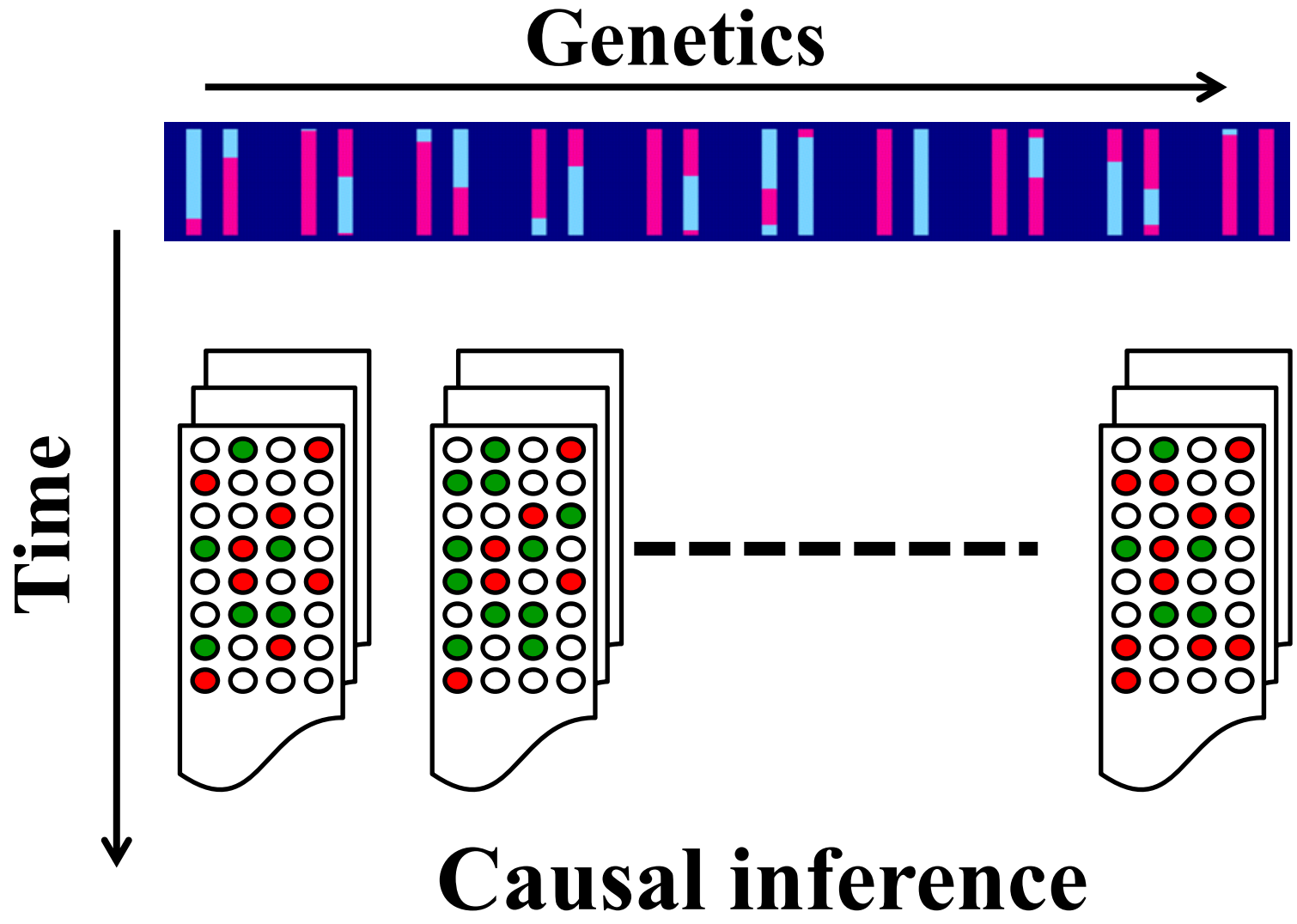
Integrating omics data into Bayesian network models

- **Genetics**
  - Zhu et al, Cytogenet Genome Res, 2004
  - Schadt et al, Nature Genetics, 2005
  - Zhu et al, PLoS Comp. Biol. 2007
- **Proteomics and Genomics**
  - Zhu et al, Nature Genetics, 2008
- **CNV**
  - Tran et al, BMC Sys. Biol, 2011
- **Metabolomics**
  - Zhu et al, PLoS Biol. 2012
- **Methylation data**
  - Yoo et al, PLoS Genetics, 2015

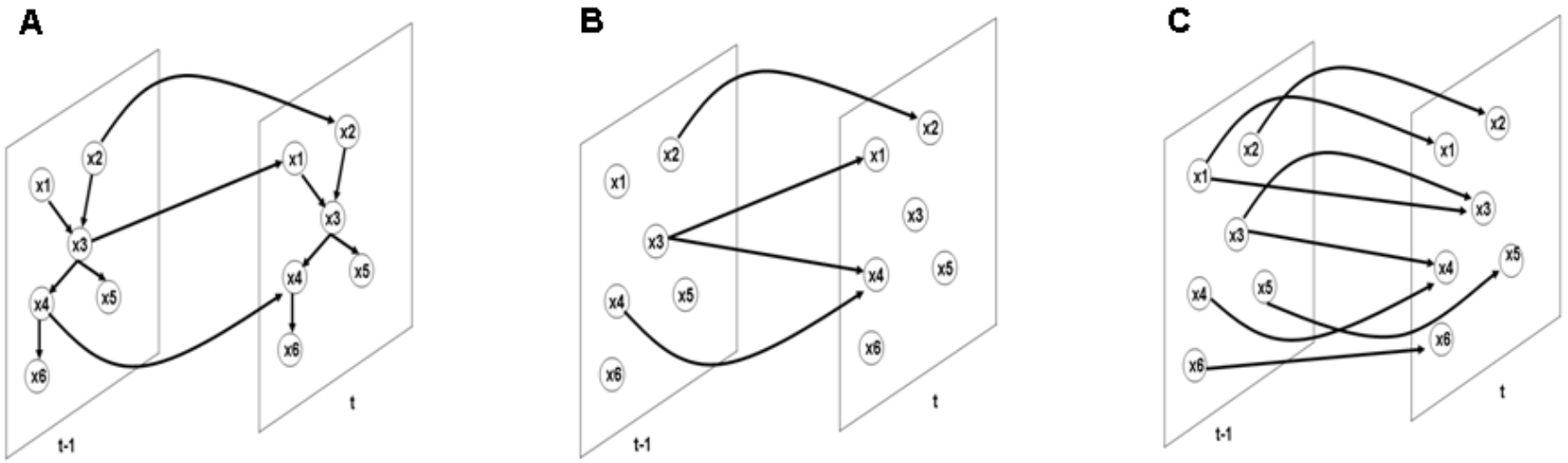
Complex traits



Integration of time dimension data into Bayesian network



Constructing causal networks

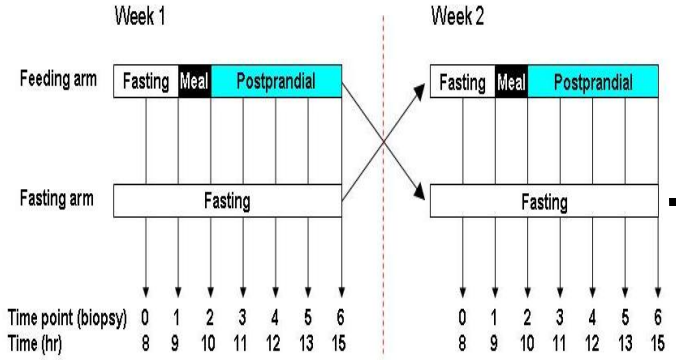


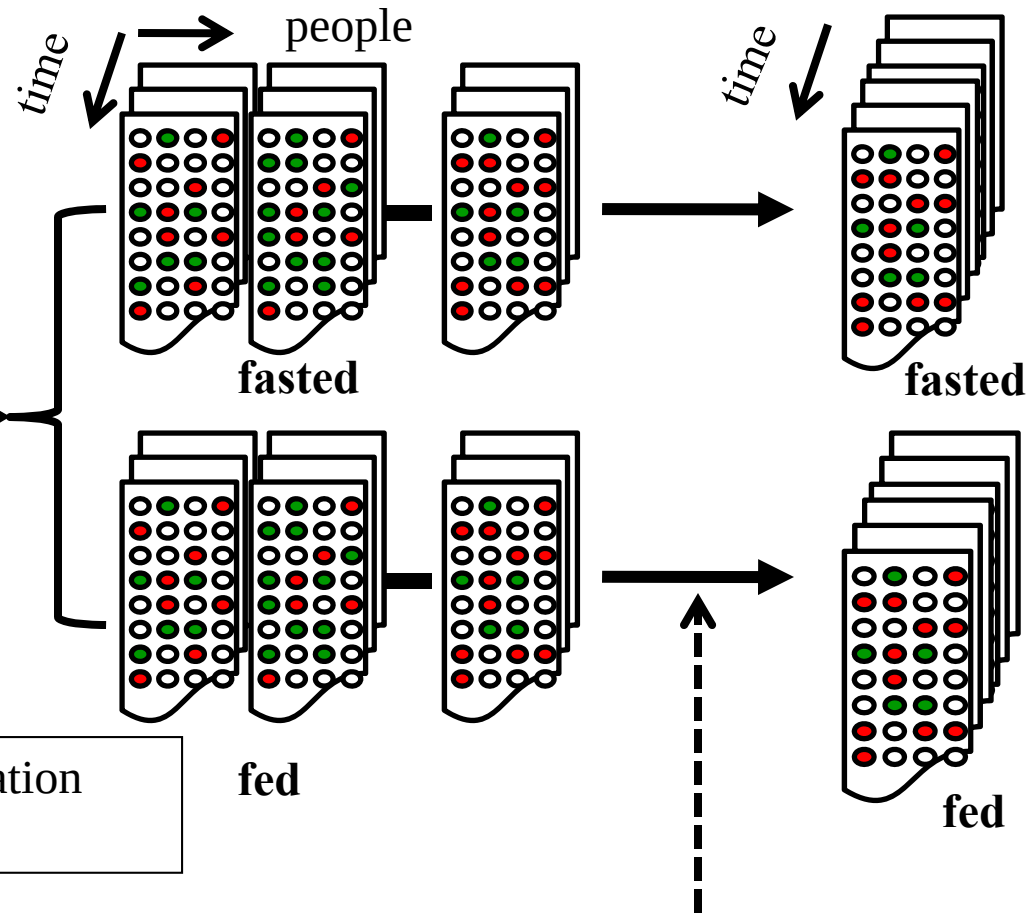
Dynamic Bayesian network

Granger causality

Zhu et al., PLoS CompBio, 2010

Human blood expression profile time series

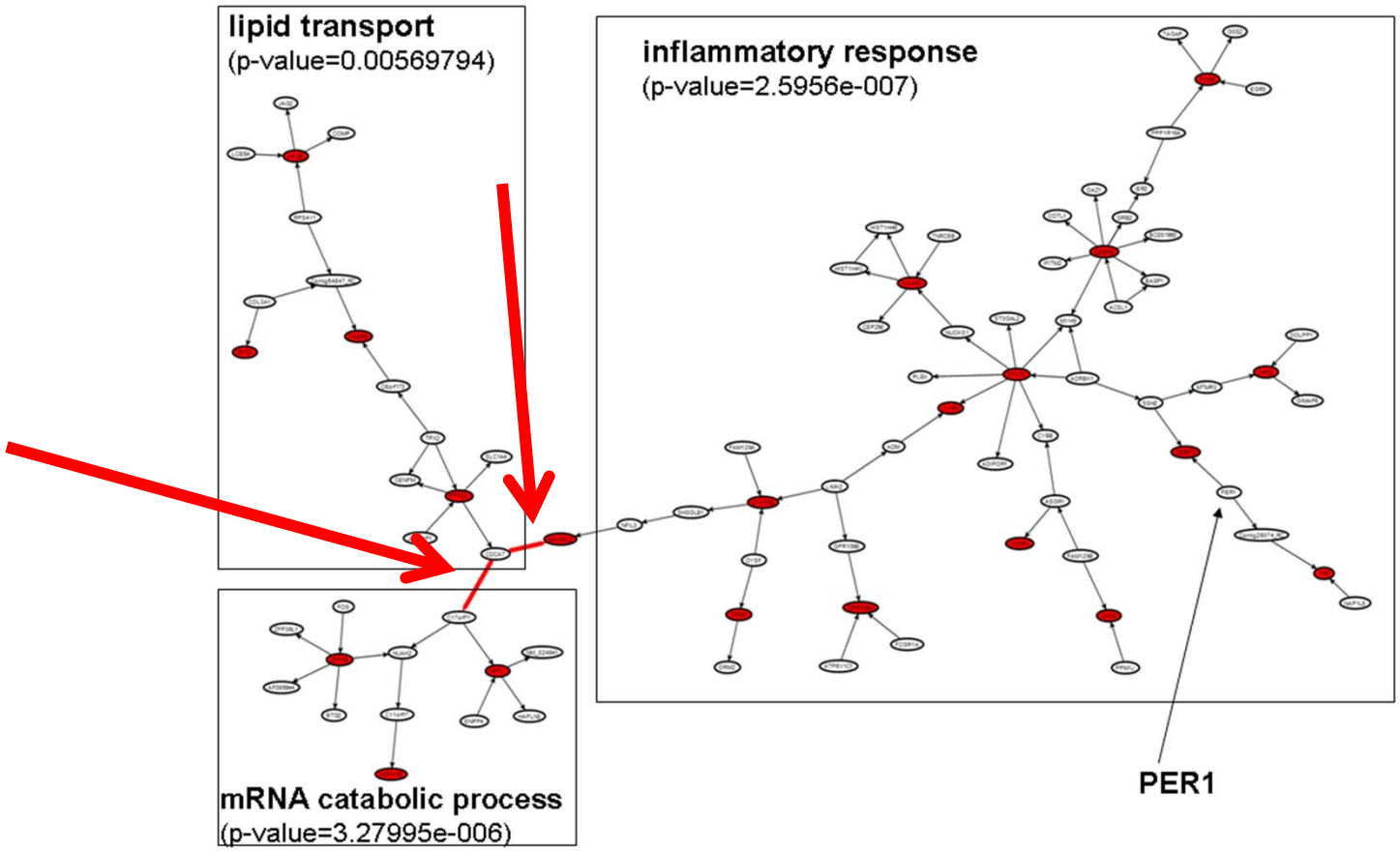




(1) Remove individual scale variation
reference time point 0

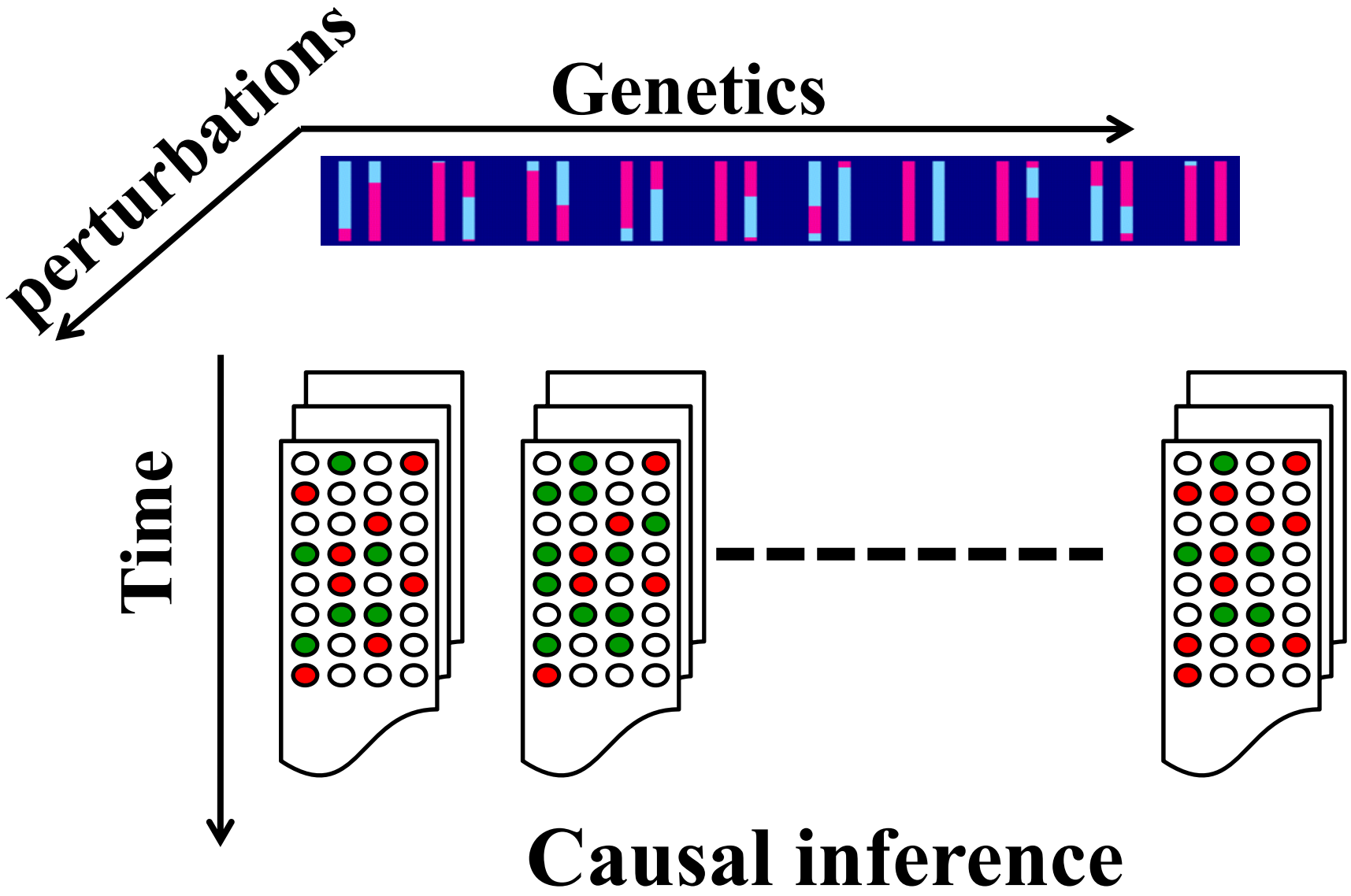
(2) Combine short time series into long ones

Integration of genetics, genomics and temporal data using Dynamic Bayesian network

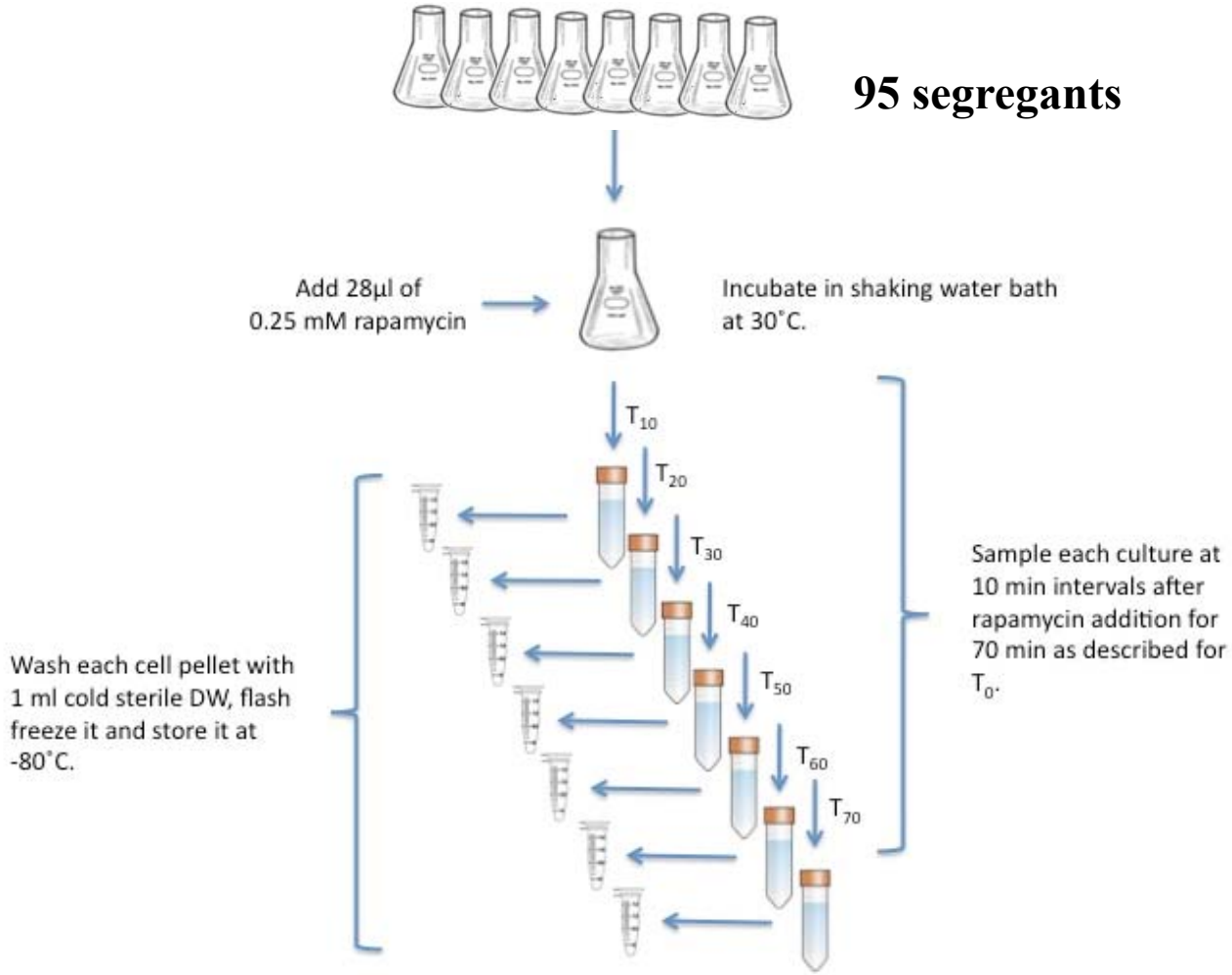


Zhu*, Chen* et al., PLoS CompBio, 2010

Integration of time dimension data into Bayesian network



Time series of drug response in a yeast F2 cross



Yeung et al., PNAS, 2011

Multi-polynomial Temporal Genetic Association

- Assume that for each genotype, time-series gene expression levels follow a multivariate normal density (Wu et al.)
- The mean vector is modeled by the cubic polynomial curve, m is the number of time points

$$\vec{g}_j = [g_j(t)]_{1\times m} = [\beta_{0j} + \beta_{1j}t + \beta_{2j}t^2 + \beta_{3j}t^3]_{1\times m}$$

- The covariance matrix is assumed to be the same for both genotyps and modeled using AR(1) repeated measurement errors (Daviddian et al; Verbeke et al)

$$\Sigma = \sigma_e^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{m-1} \\ \rho & 1 & \cdots & \rho^{m-2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{m-1} & \rho^{m-2} & \cdots & 1 \end{bmatrix}$$

Unpublished results

MPTGA

- Density for time-series data, m=6

$$f_j(y) = \frac{1}{(2\pi)^{m/2} \mid \Sigma \mid^{1/2}} \exp[(\vec{y} - \vec{g}_j)\Sigma^{-1}(\vec{y} - \vec{g}_j)^T / 2]$$

- Joint likelihood for N-95 segregants, $\Theta = (\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j}, \rho, \sigma_e^2)$

$$L(\Theta) = \prod_{i=1}^{N} [\delta_{i0} f_0(\vec{y}_i) + \delta_{i1} f_1(\vec{y}_i)]$$

- Maximum likelihood estimate (MLE)
- Likelihood ratio test

$$H_0 : \beta_{00} = \beta_{01}, \beta_{10} = \beta_{11}, \beta_{20} = \beta_{21}, \beta_{30} = \beta_{31}$$

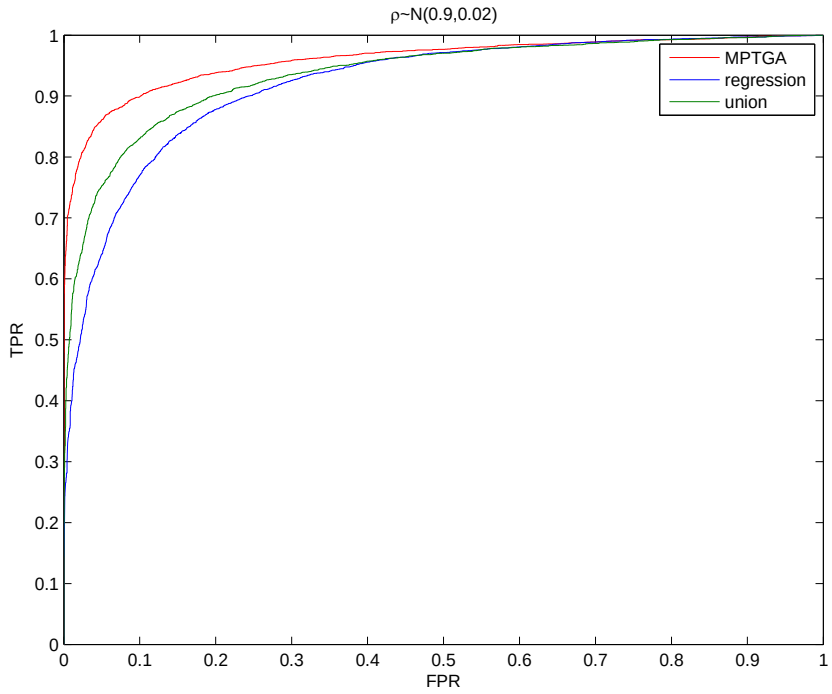
$$H_1 : \text{at least one of the equalities does not hold}$$

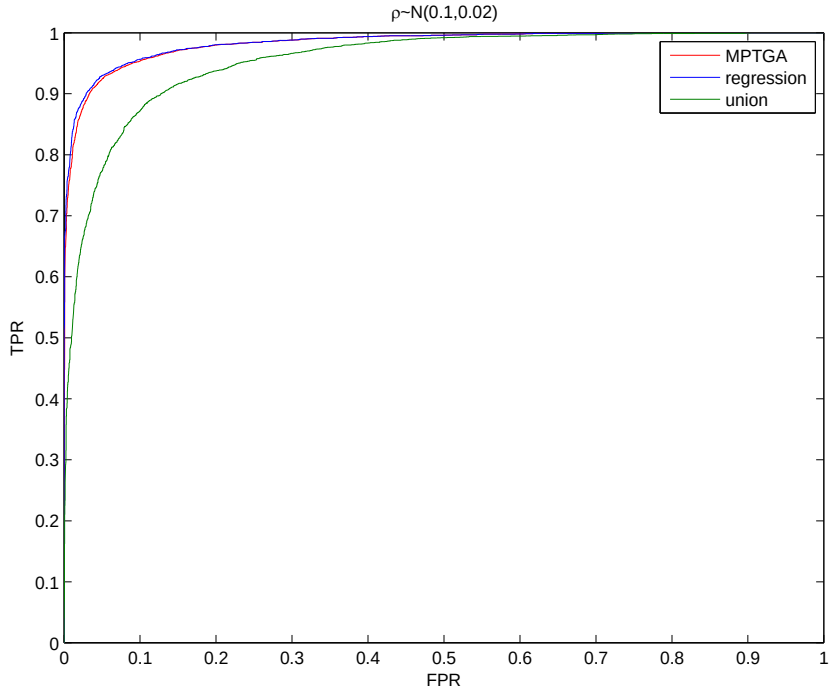
Simulation Study

- Simulate time series data from multivariate normal distribution, with mean vector modeled by various patterns that are similar to the observed experimental results.
- Data drawn either from a single model or two separate models
- 10,000 groups of time series data were simulated, in which each group is composed of 100 6-point time series data
- Performance comparison
  - ROC curve
  - simulation under different auto-regressive coefficient ρ

ROC curves

- ROC curve comparing performance among three different time-dependent genetic association methods





Unpublished results

Mathematical Models

eQTL effect

M1: $X_{i,t} = \delta_{i0}(\alpha_{00} + \alpha_{10}X_{i,t-1}) + \delta_{i1}(\alpha_{01} + \alpha_{11}X_{i,t-1}) + \varepsilon_i$

$Y_{i,t} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_2 X_{i,t-1} + \mu_i$

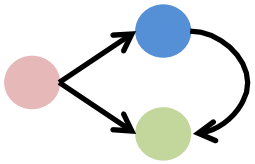
Auto-regressive term

Causal effect

M2: $Y_{i,t} = \delta_{i0}(\beta_{00} + \beta_{10}Y_{i,t-1}) + \delta_{i1}(\beta_{01} + \beta_{11}Y_{i,t-1}) + \mu_i$

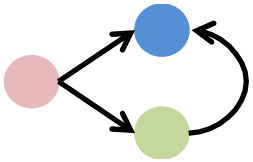
$X_{i,t} = \alpha_0 + \alpha_1 X_{i,t-1} + \alpha_2 Y_{i,t-1} + \varepsilon_i$

M3: $X_{i,t} = \delta_{i0}(\alpha_{00} + \alpha_{10}X_{i,t-1}) + \delta_{i1}(\alpha_{01} + \alpha_{11}X_{i,t-1}) + \varepsilon_i$



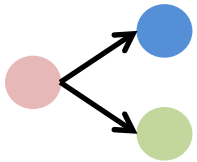
$Y_{i,t} = \delta_{i0}(\beta_{00} + \beta_{10}Y_{i,t-1}) + \delta_{i1}(\beta_{01} + \beta_{11}Y_{i,t-1}) + \beta_2 X_{i,t-1} + \mu_i$

M4: $Y_{i,t} = \delta_{i0}(\beta_{00} + \beta_{10}Y_{i,t-1}) + \delta_{i1}(\beta_{01} + \beta_{11}Y_{i,t-1}) + \mu_i$



$X_{i,t} = \delta_{i0}(\alpha_{00} + \alpha_{10}X_{i,t-1}) + \delta_{i1}(\alpha_{01} + \alpha_{11}X_{i,t-1}) + \alpha_2 Y_{i,t-1} + \varepsilon_i$

M5: $X_{i,t} = \delta_{i0}(\alpha_{00} + \alpha_{10}X_{i,t-1}) + \delta_{i1}(\alpha_{01} + \alpha_{11}X_{i,t-1}) + \varepsilon_i$



$Y_{i,t} = \delta_{i0}(\beta_{00} + \beta_{10}Y_{i,t-1}) + \delta_{i1}(\beta_{01} + \beta_{11}Y_{i,t-1}) + \mu_i$

Model Selection: BIC or AIC

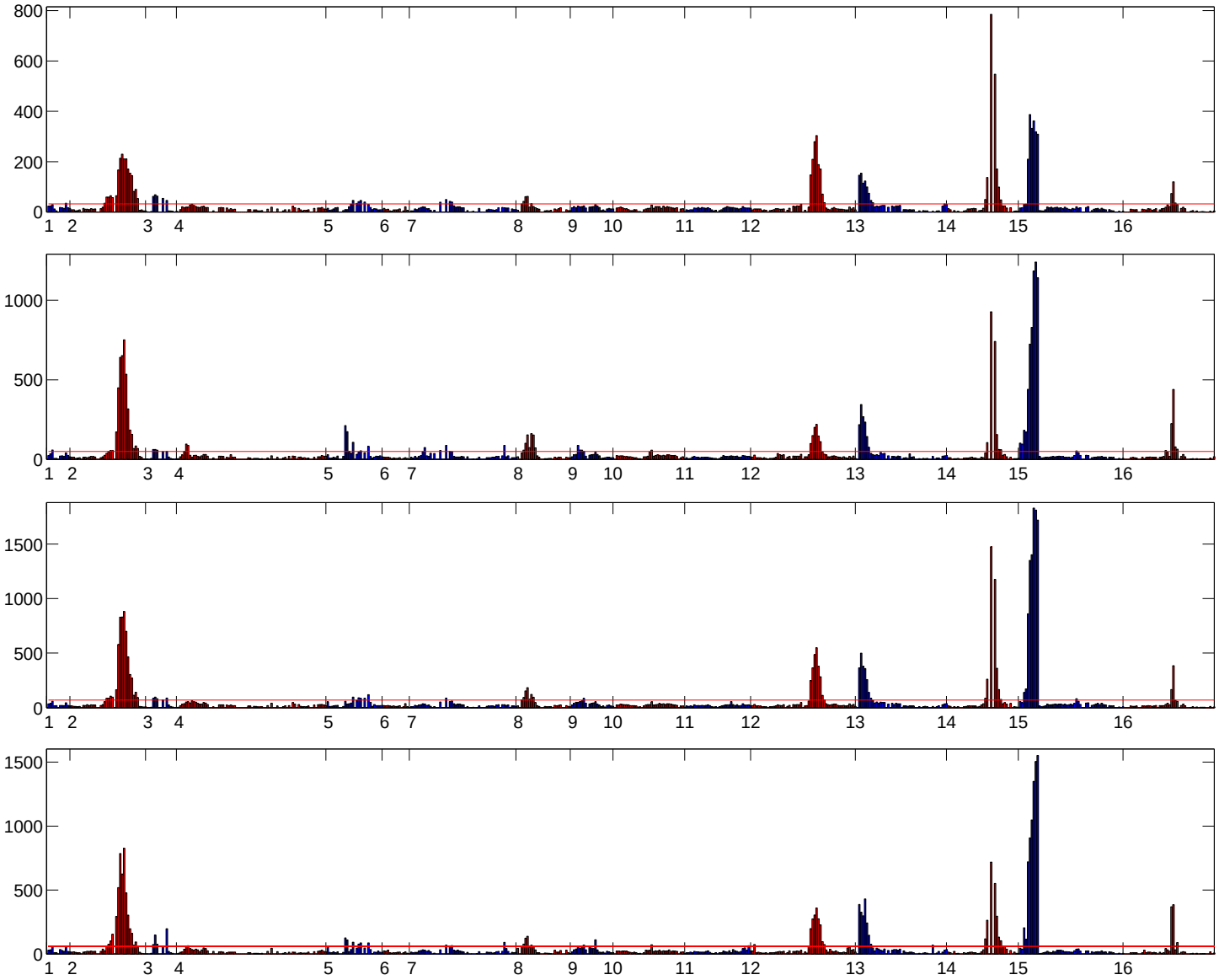
Unpublished results

eQTL and CI

	Static (T0)	MPTGA	Union	Regression
# eQTL identified	2698	3078	5336	2534
Avg. 1-LOD drop CI (kb)	41.44	16.91	46.99	7.37

Histogram plot

- Number of linkages plotted against genome location by static approach and three time-dependent approaches.



Hotspots identified and enrichment analysis of Rapamycin respond signature

HotSpot	Chr	Pos	Size				Overlap with Rapamycin Signiture						
			T0(>=51)	MP(>=52)	UNION(>=93)	REG(>=82)	T0	MP	MP-T0	UNION	UNION-T0	REG	REG-T0
1	5	10,000	63	28	89	80	0.791445	N/A	N/A	N/A	N/A	N/A	N/A
2	6	230,000	60	10	57	25	1.000000	N/A	N/A	N/A	N/A	N/A	N/A
3	10	390,000	53	57	88	140	1.000000	0.263320	0.093846	N/A	N/A	0.190718	0.052744
4	3	210,000	82	49	119	307	0.521534	N/A	N/A	0.641771	1.000000	0.165317	0.235254
5	2	550,000	309	737	1057	1044	0.340028	0.010014	0.074378	0.902415	0.959815	0.383476	0.598213
6	3	90,000	95	63	121	211	0.028996	0.000506	0.048672	0.049577	1.000000	0.102558	0.658762
7	5	430,000	59	81	164	143	0.760136	0.724516	1.000000	0.630408	0.603578	0.483478	0.587675
8	7	370,000	60	84	115	100	1.000000	1.000000	1.000000	0.903533	0.522707	0.485801	0.105384
9	8	110,000	101	151	237	172	0.315836	0.253973	0.143157	0.651793	0.913613	0.679832	0.693857
10	12	670,000	425	215	666	451	0.989515	0.657945	0.724856	0.999955	0.999073	0.997853	1.000000
11	13	50,000	260	331	660	455	0.007584	0.004479	0.239248	0.009641	0.243907	0.000722	0.035084
12	14	450,000	1030	906	1692	899	0.822970	0.559392	0.299970	0.917223	0.865509	0.671883	0.391116
13	14	490,000	751	722	1396	704	0.818425	0.532639	0.578993	0.955582	0.932882	0.622181	0.475578
14	15	110,000	557	709	1611	1148	0.143048	8.92553E-07	0.001653	0.002092	0.010411	0.001866	0.017717
15	15	170,000	450	1218	2043	1724	0.007575	1.08694E-05	0.006387	0.009224	0.180087	0.000711	0.047979
16	16	510,000	254	425	575	555	0.620547	0.000410	0.002262	0.474326	0.328784	0.319300	0.445220
17	1	50000	45	56	96	100	N/A	0.003634	0.001445	0.939067	1.000000	0.172711	0.152071
18	4	90,000	50	92	96	101	N/A	6.34349E-05	6.68703E-06	0.823102	0.751701	0.686390	0.511770
19	5	190,000	15	207	140	234	N/A	1.56567E-10	6.95272E-11	0.011356	0.005820	0.002211	0.001210
20	7	150,000	37	72	46	42	N/A	1.41785E-08	4.97133E-10	N/A	N/A	N/A	N/A
21	9	70,000	34	84	75	81	N/A	9.71346E-05	9.39187E-06	N/A	N/A	N/A	N/A
22	9	250,000	50	43	78	181	N/A	N/A	N/A	N/A	N/A	0.448413	0.380269
23	12	30,000	32	25	49	172	N/A	N/A	N/A	N/A	N/A	0.094553	0.094009
24	13	790,000	10	15	30	149	N/A	N/A	N/A	N/A	N/A	0.018386	0.014940
25	7	970,000	23	87	33	137	N/A	0.000140	2.10321E-05	N/A	N/A	0.000150	9.93968E-05
26	8	130,000	33	73	102	71	N/A	0.115927	0.223108	0.324007	0.410008	N/A	N/A
			16	20	18	22	3	13		4		7	

GO term enrichment analysis for eQTL hot spots identified by MPTGA

HotSpot	Chr	Pos	GO-Term	p-value
1	5	10,000	cytosolic large ribosomal subunit (sensu Eukaryota)	2.31E-06
2	6	230,000	N/A	
3	10	390,000	N/A	
4	3	210,000	N/A	
5	2	550,000	structural constituent of ribosome	3.25E-43
6	3	90,000	branched chain family amino acid biosynthesis	2.89E-09
7	5	430,000	N/A	
8	7	370,000	protein folding	1.12E-14
9	8	110,000	mating projection tip	8.40E-08
10	12	670,000	ergosterol biosynthesis	6.29E-25
11	13	50,000	phosphate transport	2.15E-07
12	14	450,000	endopeptidase activity	5.55E-16
13	14	490,000	endopeptidase activity	5.66E-20
14	15	110,000	structural constituent of ribosome	2.48E-41
15	15	170,000	structural constituent of ribosome	1.45E-35
16	16	510,000	structural constituent of ribosome	1.55E-54
17	1	50000	N/A	
18	4	90,000	structural constituent of ribosome	3.21E-28
19	5	190,000	structural constituent of ribosome	1.67E-102
20	7	150,000	structural constituent of ribosome	5.58E-36
21	9	70,000	structural constituent of ribosome	5.50E-37
22	9	250,000	N/A	
23	12	30,000	N/A	
24	13	790,000	N/A	
25	7	970,000	structural constituent of ribosome	1.06E-51
26	8	130,000	mating projection tip	1.10E-07

- A key mechanism of cell controlling growth is to regulate ribosome biogenesis.
- Ribosomal protein gene expression is regulated by mTOR, target of rapamycin
- All eQTL hot spots identified by MPTGA method and enriched for the rapamycin respond signature are enriched for GO term structural constituent of ribosome

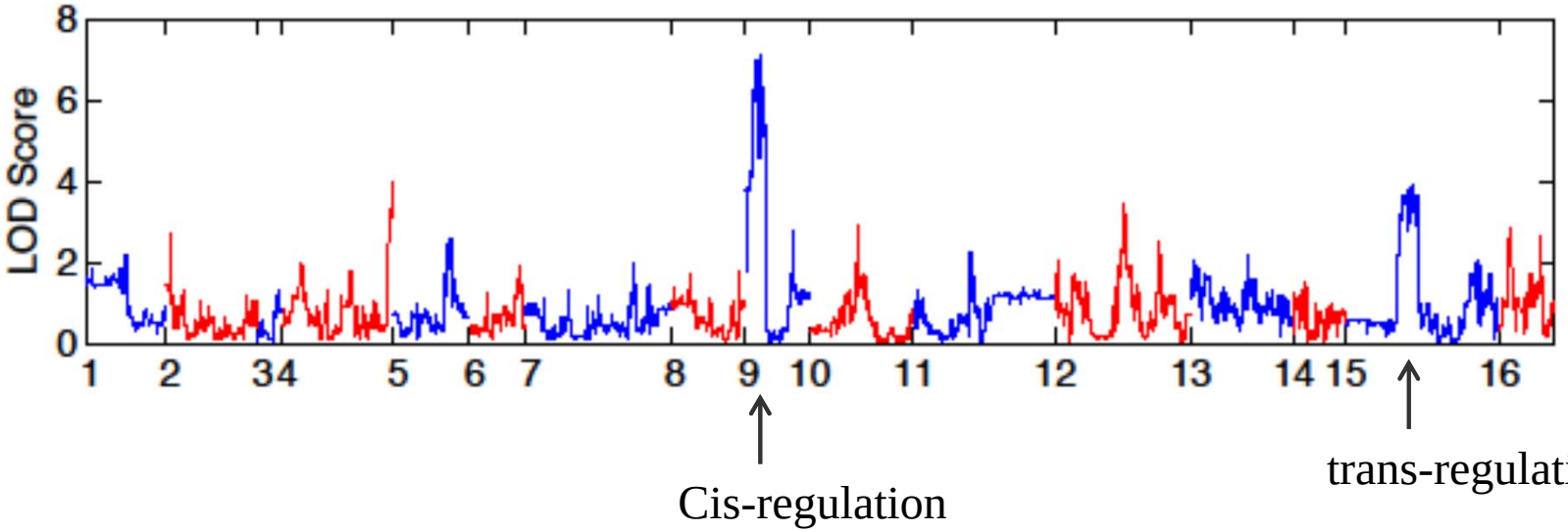
Causal regulator identification

- Causal regulators identification by time-dependent genetic causality test, and comparison with previous findings in Yvert et al and Zhu et al. Genes in red bold font are overlapping with previous findings.

HotSpot	Chr	Pos	Yvert et al.	BN Full	TGCT(BIC)
1	5	10,000	N/A	N/A	N/A
2	6	230,000	N/A	N/A	N/A
3	10	390,000	N/A	N/A	YR016W
4	3	210,000	**METALPHA1**	**METALPHA1**	**METALPHA1**
5	2	550,000	AMN1, MAK5	TBS1, TOS1,ARA1, CSH1, SUP45, CNS1, AMN1	SEC66, ICS2, RPBS, UBS1
6	3	90,000	**LEU2**	**LEU2**, ILV6, NFS1, CIT2, METALPHAL1	**LEU2**
7	5	430,000	N/A	N/A	AVT6
8	7	370,000	N/A	N/A	MRH4, YGL081W, PUS2
9	8	110,000	N/A	N/A	ERG11, SHU1
10	12	670,000	**HAP1**	**HAP1**	**HAP1**
11	13	50,000	None	None	MDM1
12	14	450,000	N/A	N/A	PHO23, TOP2, MKT1, YPT53
13	14	490,000	None	SAL1, TOP2	TOP2, YDJ1, **MKT1**
14	15	110,000	N/A	N/A	MKT1
15	15	170,000	None	**PHM7**	YOL073C, ATP19, **PHM7**, MSH2, HAL9
16	16	510,000	N/A	N/A	YPL038W-A, MET12, MET31, RMI1, YPL039W
17	1	50000	N/A	N/A	PUS2, ASF2
18	4	90,000	N/A	N/A	YDL203C, ASF2, PRR2, RTN2
19	5	190,000	N/A	N/A	ISC1, SMB1
20	7	150,000	N/A	N/A	UBS1, YLC1353W
21	9	70,000	N/A	N/A	RRD1, ECM37, YL130W
22	9	250,000	N/A	N/A	N/A
23	12	30,000	N/A	N/A	N/A
24	13	790,000	N/A	N/A	N/A
27	7	970,000	N/A	N/A	CPD1, YHB1, SOL4
28	8	130,000	**GPA1**	**GPA1**	ERG11, **GPA1**

RRD1 as a causal gene for the eQTL hot spot at chrIX:70,000

RRD1 eQTL based on MPTGA



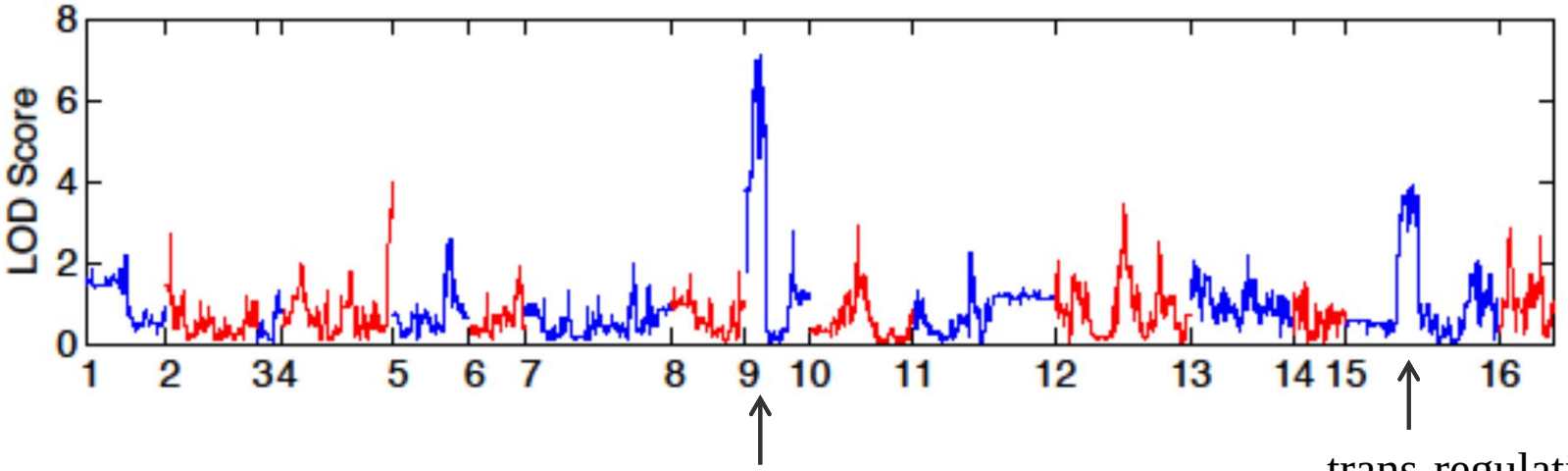
RRD1 ko vs wildtype

- **65 genes differentially expressed**
- **3 of them overlap with genes in eQTL hot spot chrIX:70,000**
- **21 of them overlap with genes in eQTL hot spot chrXV:170,000 (5.7 fold enrichment, p-value=1.9e-11)**

Lin et al., (unpublished)

RRD1 as a causal gene for the eQTL hot spot at chrIX:70,000

RRD1 eQTL based on MPTGA

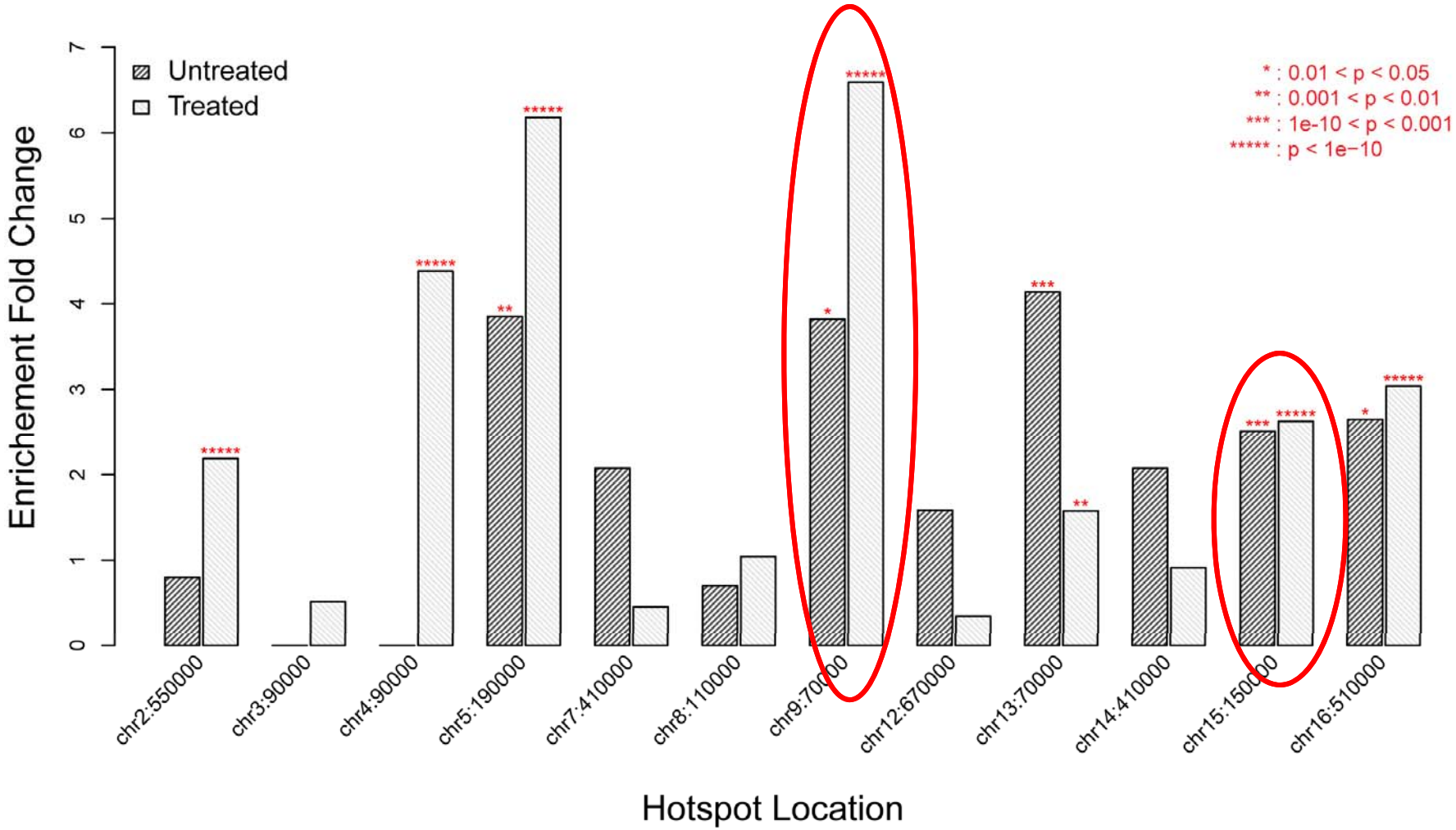


RRD1 ko vs wildtype both treated with rapamycin

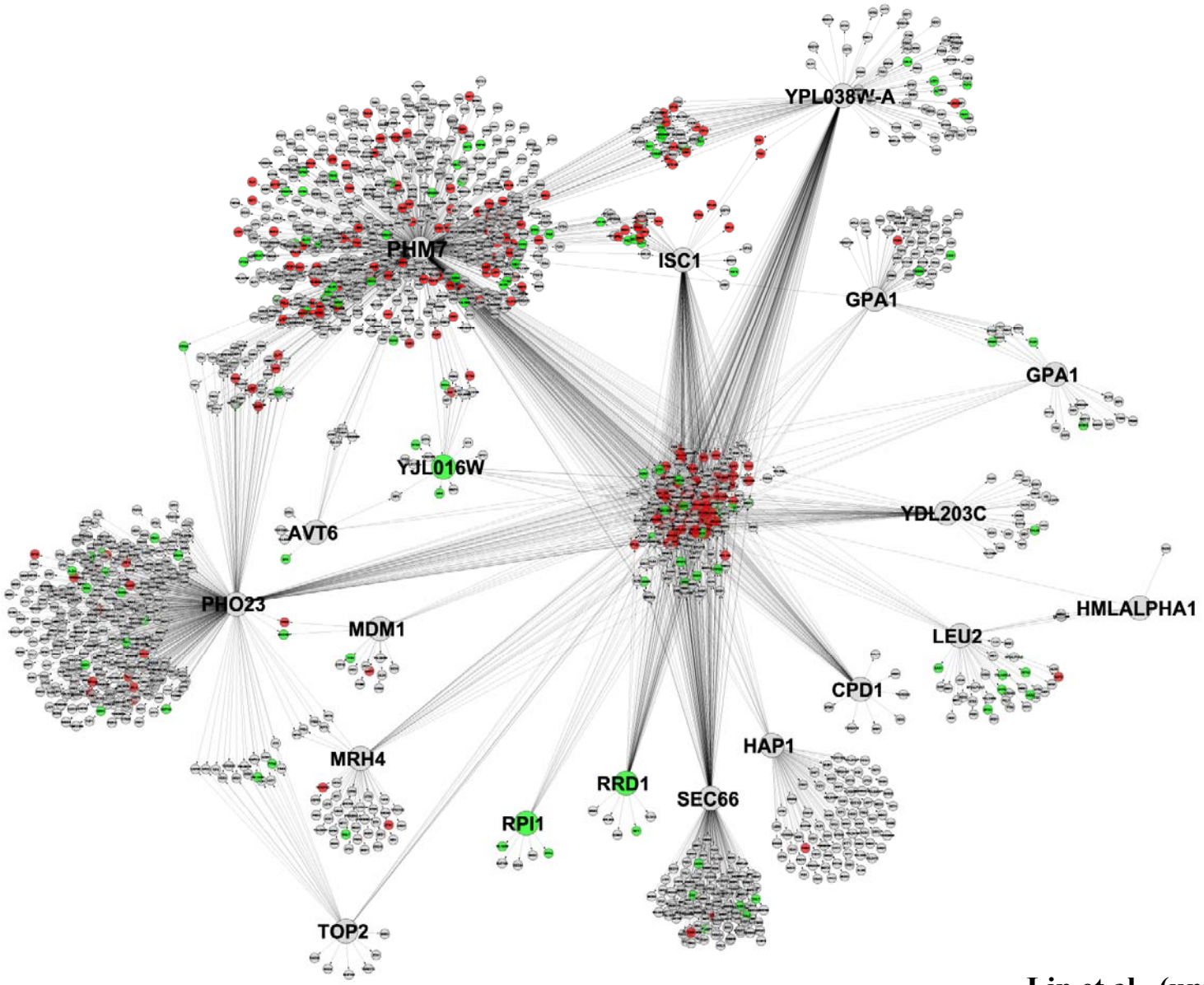
- **584 genes differentially expressed**
- **52 of them overlap with genes in eQTL hot spot chrIX:70,000 (6.1 fold enrichment, p=2.0e-31)**
- **35 of them overlap with genes in eQTL hot spot chrXV:170,000 (2.5 fold enrichment, p-value=5.6e-79)**

Lin et al., (unpublished)

RRD1 as a causal gene for the eQTL hot spot at chrIX:70,000



Lin et al., (unpublished)



Lin et al., (unpublished)

Aknowledgements

Zhu lab
Seungyeul Yoo
Eunjee Lee
Li Wang
Luan Lin
Quan Long

Mount Sinai Genomics Institute
Eric Schadt
Bin Zhang
Zhidong Tu
Charles Powell
Patrizia Casaccia

Boston University
Avrum Spira
Joshua Campbell

U Washington
Roger Baumgarner

Berkerley
Rachel Brem

Princeton
Lenoid Kruglyak

Supported by:

•Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai
•Janssen
•Canary Foundation
•Prostate Cancer Foundation
•NIH
•NCI