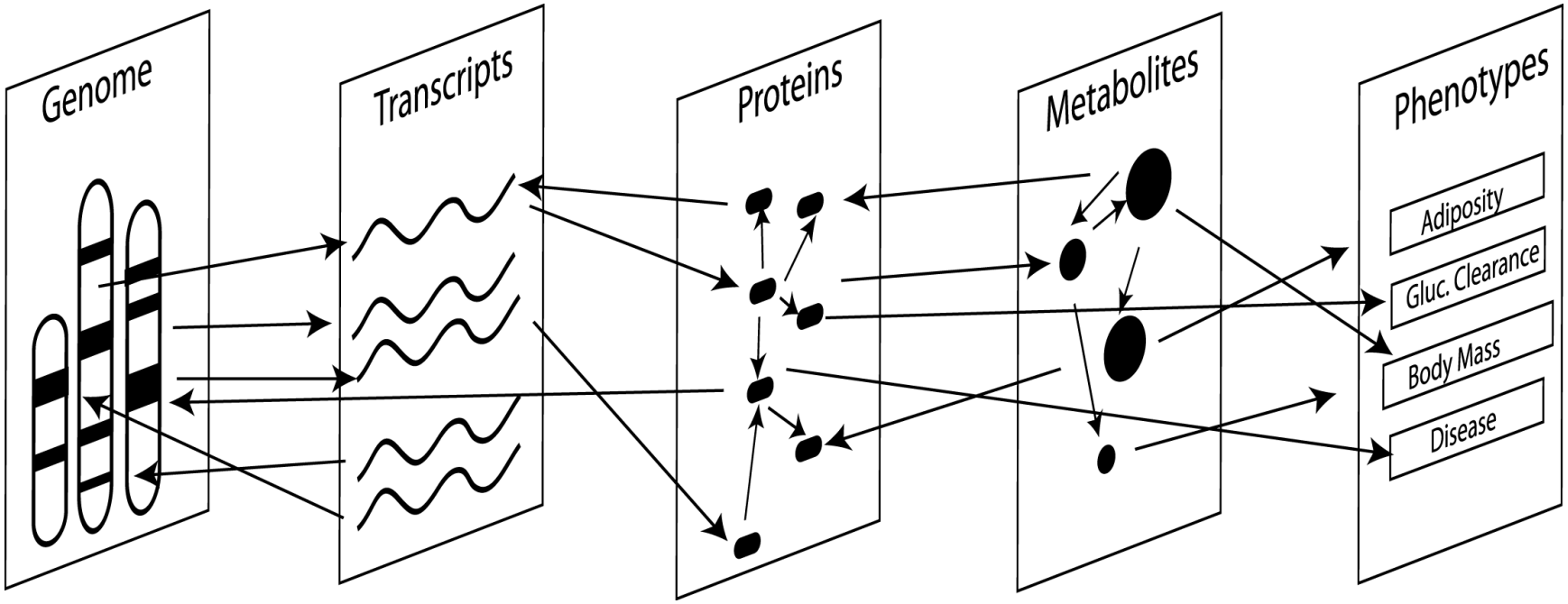
Belief Propagation in Genotype-Phenotype Networks

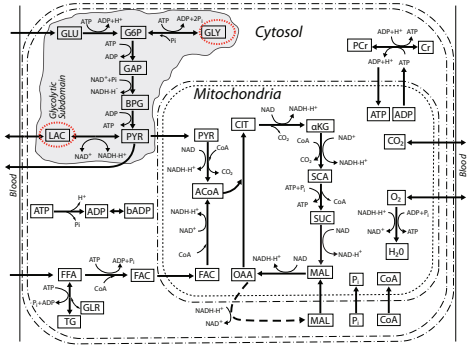
Rachael Hageman Blair

Systems Biology



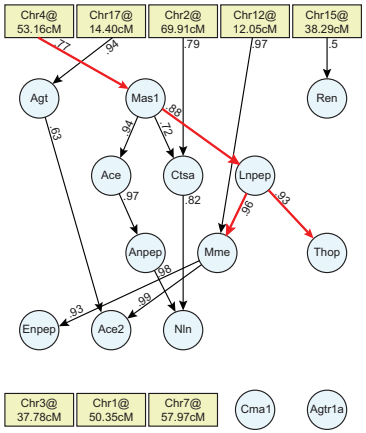
“If you want truly to understand something, try to change it”
-Kurt Lewin 1947 (Social psychology pioneer)

Outline



I. Deterministic Metabolic Models

- Model Development
- The Inverse Problem



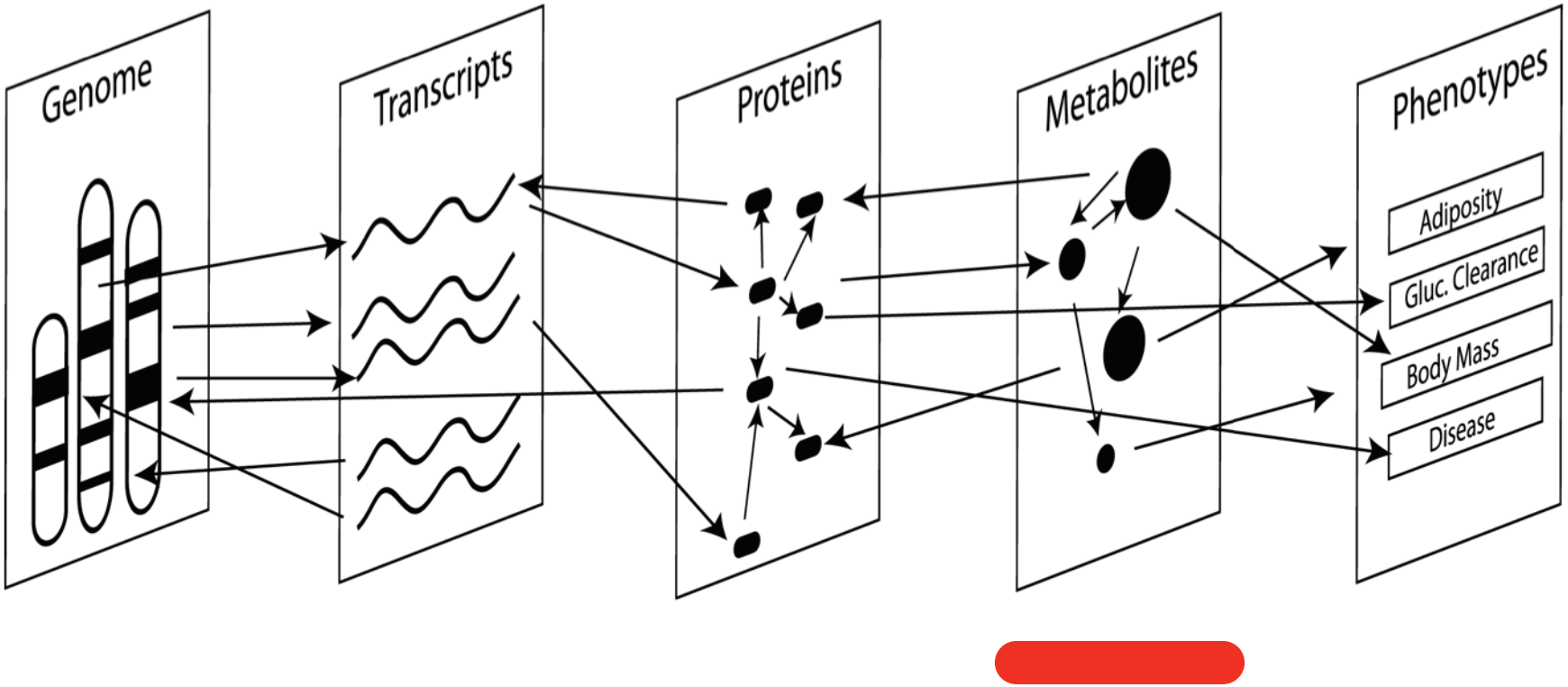
II. Causal Graphical Models

- Preliminaries: data, QTL
- Approaches and Limitations

III. Belief Propagation in Genotype-Phenotype Networks

- The modeling
- Prediction
- Stability

I. Deterministic Metabolic Models



Modeling Metabolic Systems

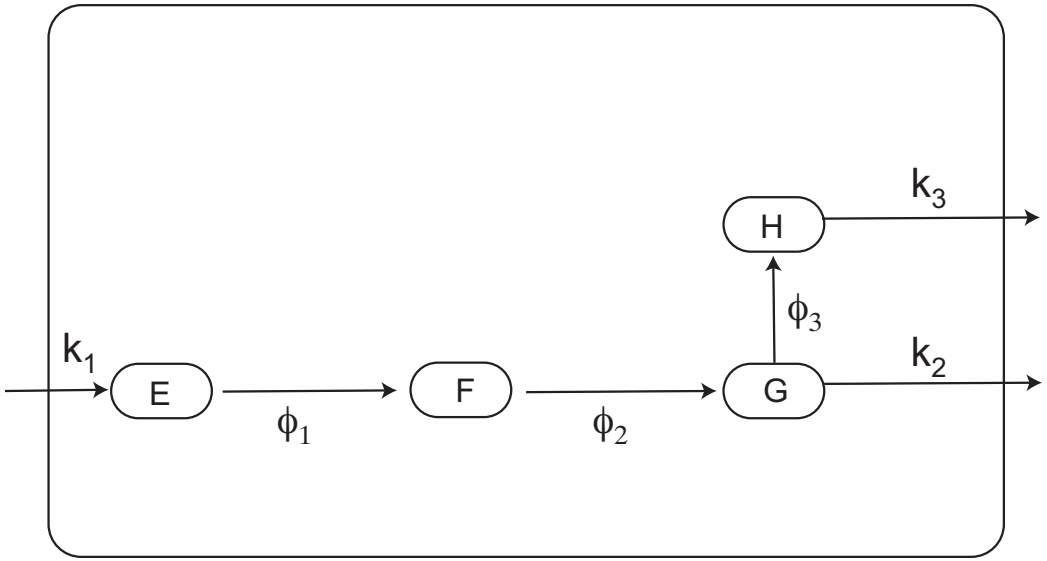
$$\frac{dC}{dt} = \text{Production} - \text{Utilization}$$

$$\frac{dE}{dt} = k_1 - \phi_1$$

$$\frac{dF}{dt} = \phi_1 - \phi_2$$

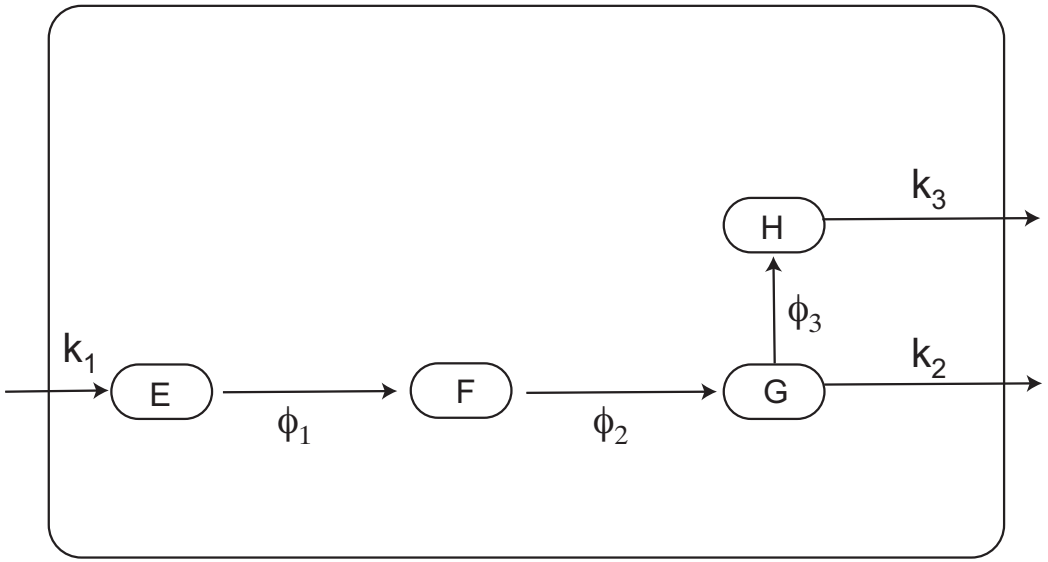
$$\frac{dG}{dt} = \phi_2 - \phi_3 - k_2$$

$$\frac{dH}{dt} = \phi_3 - k_3$$



Parameter Estimation

$$\phi_1 : A \rightarrow B, \quad \phi_1 = V_{\max,1} \frac{C_A}{K_1 + C_A}$$



$$\frac{dE}{dt} = k_1 - V_{\max,1} \frac{C_E}{K_1 + C_E}$$

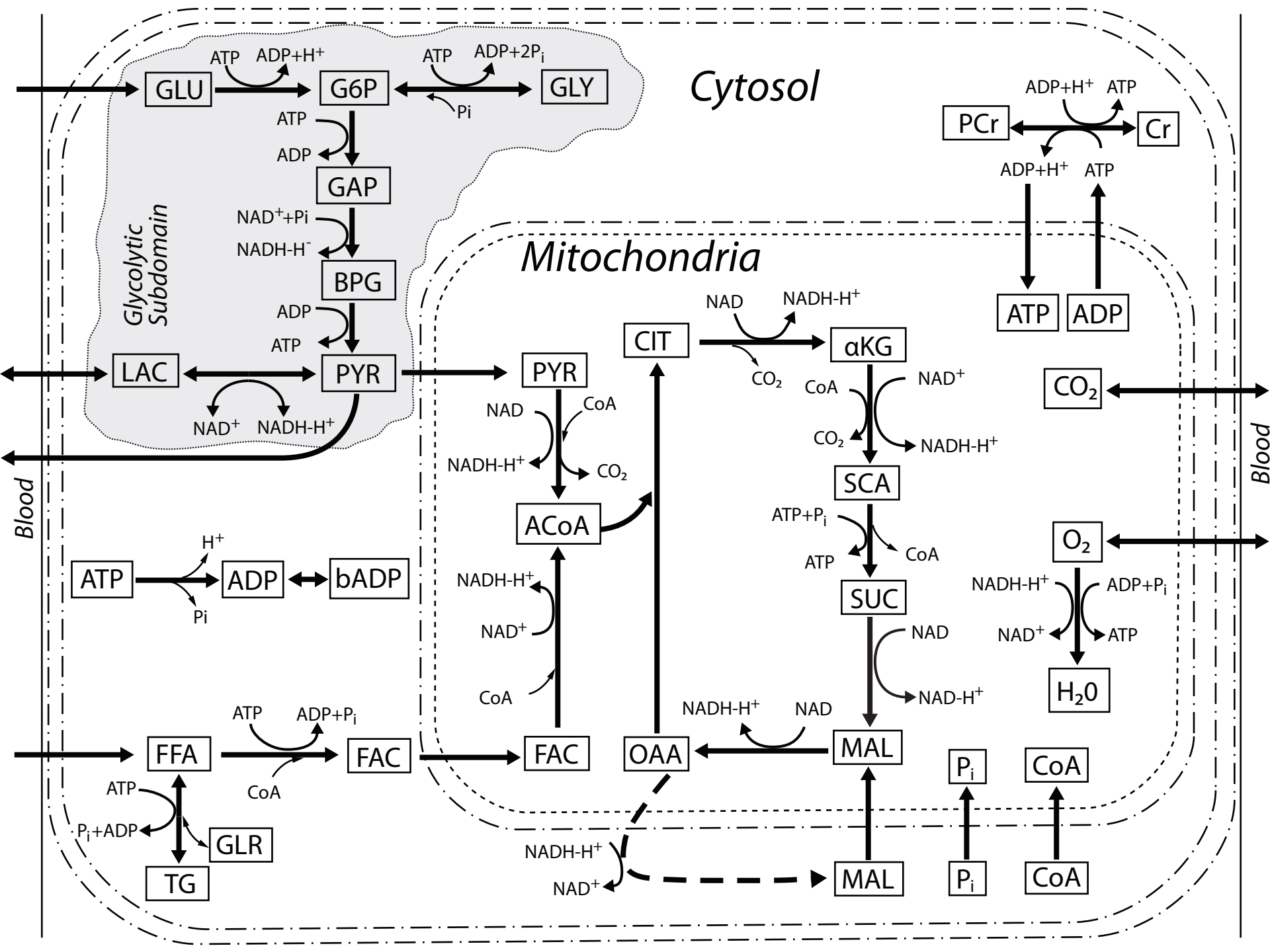
$$\frac{dF}{dt} = V_{\max,1} \frac{C_E}{K_1 + C_E} - V_{\max,2} \frac{C_F}{K_2 + C_F}$$

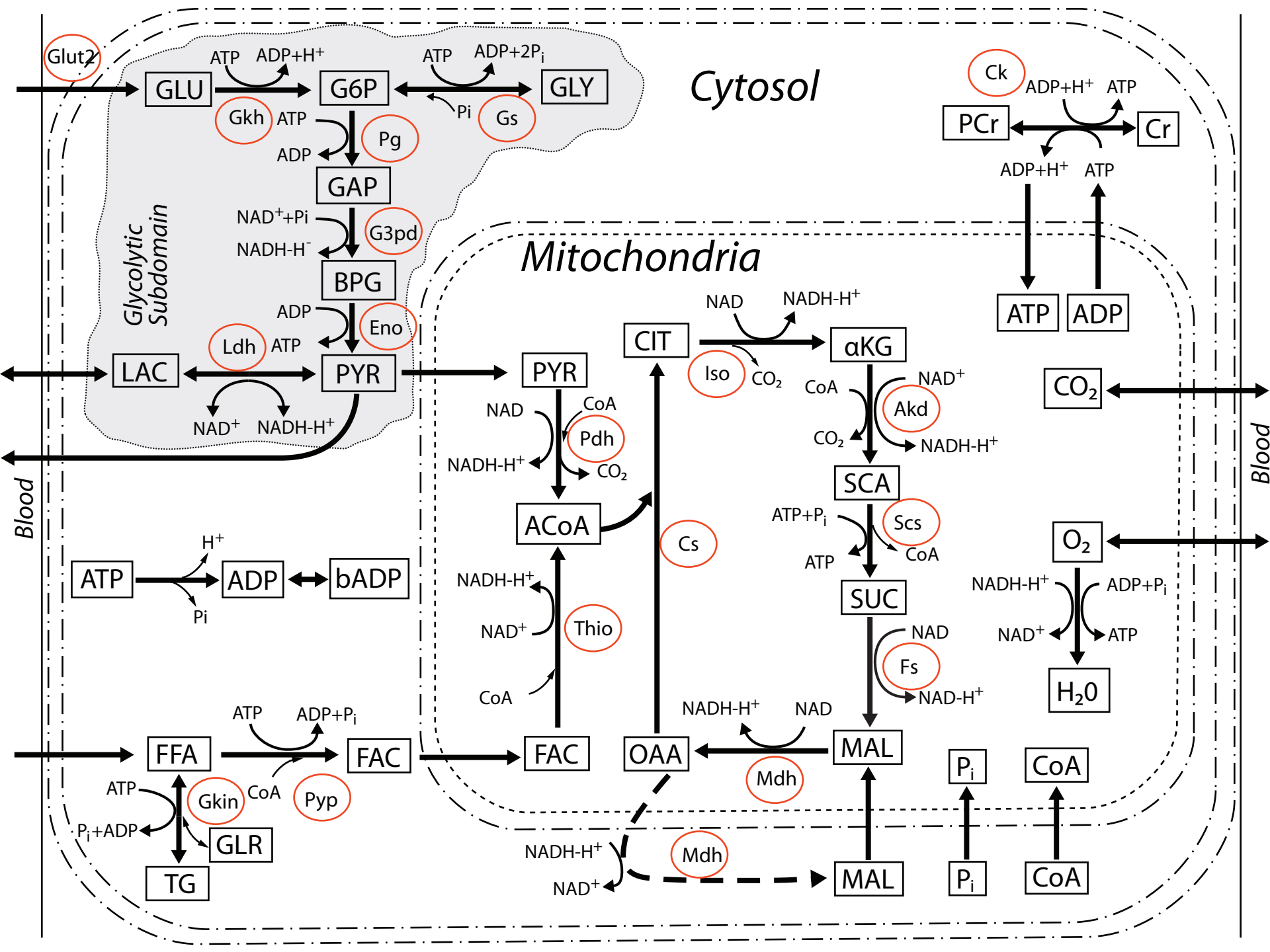
$$\frac{dG}{dt} = V_{\max,2} \frac{C_F}{K_2 + C_F} - V_{\max,3} \frac{C_G}{K_3 + C_G} - k_2$$

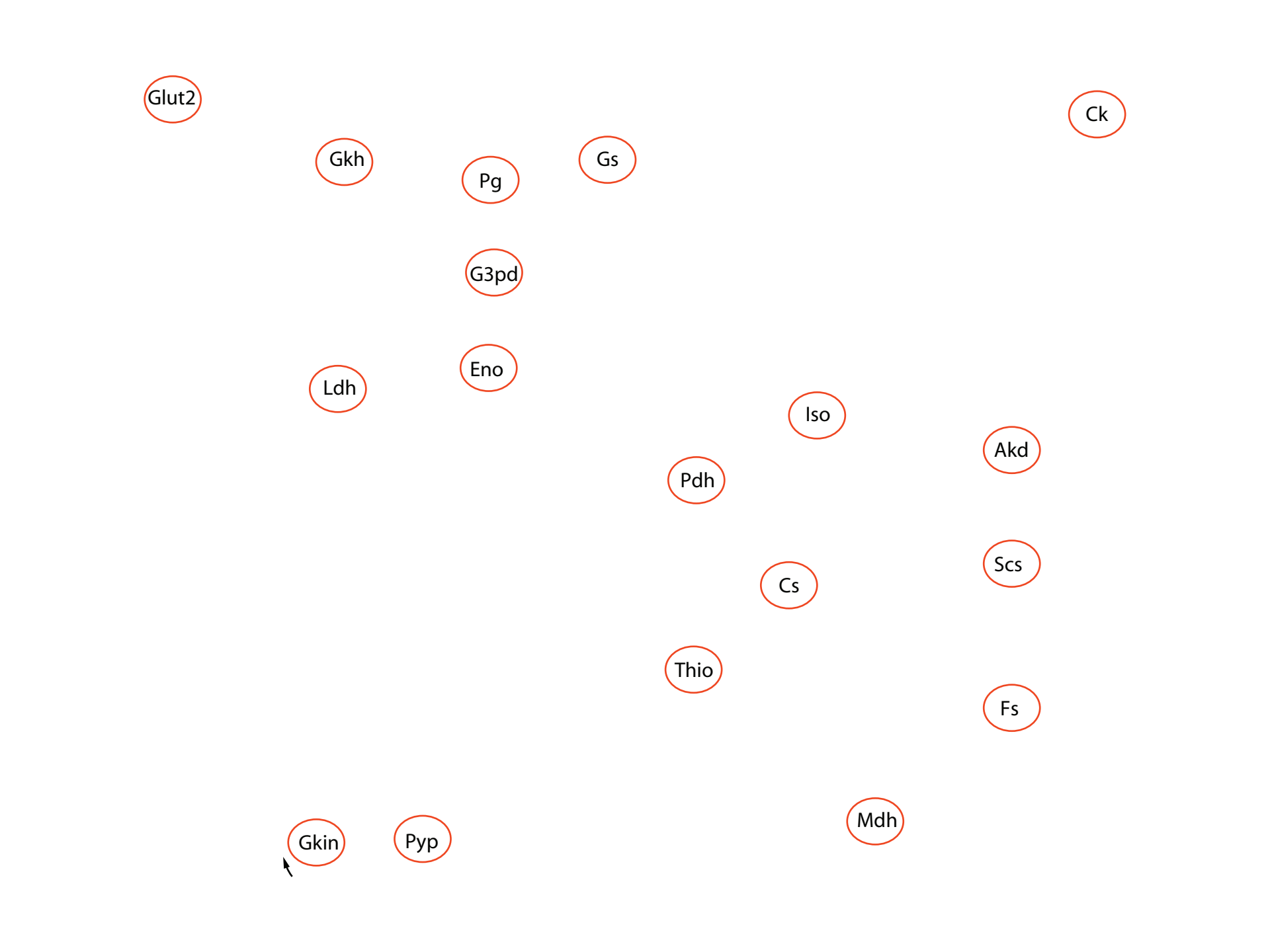
$$\frac{dH}{dt} = V_{\max,3} \frac{C_G}{K_3 + C_G} - k_3$$

Underdetermined Parameter Estimation Problem!

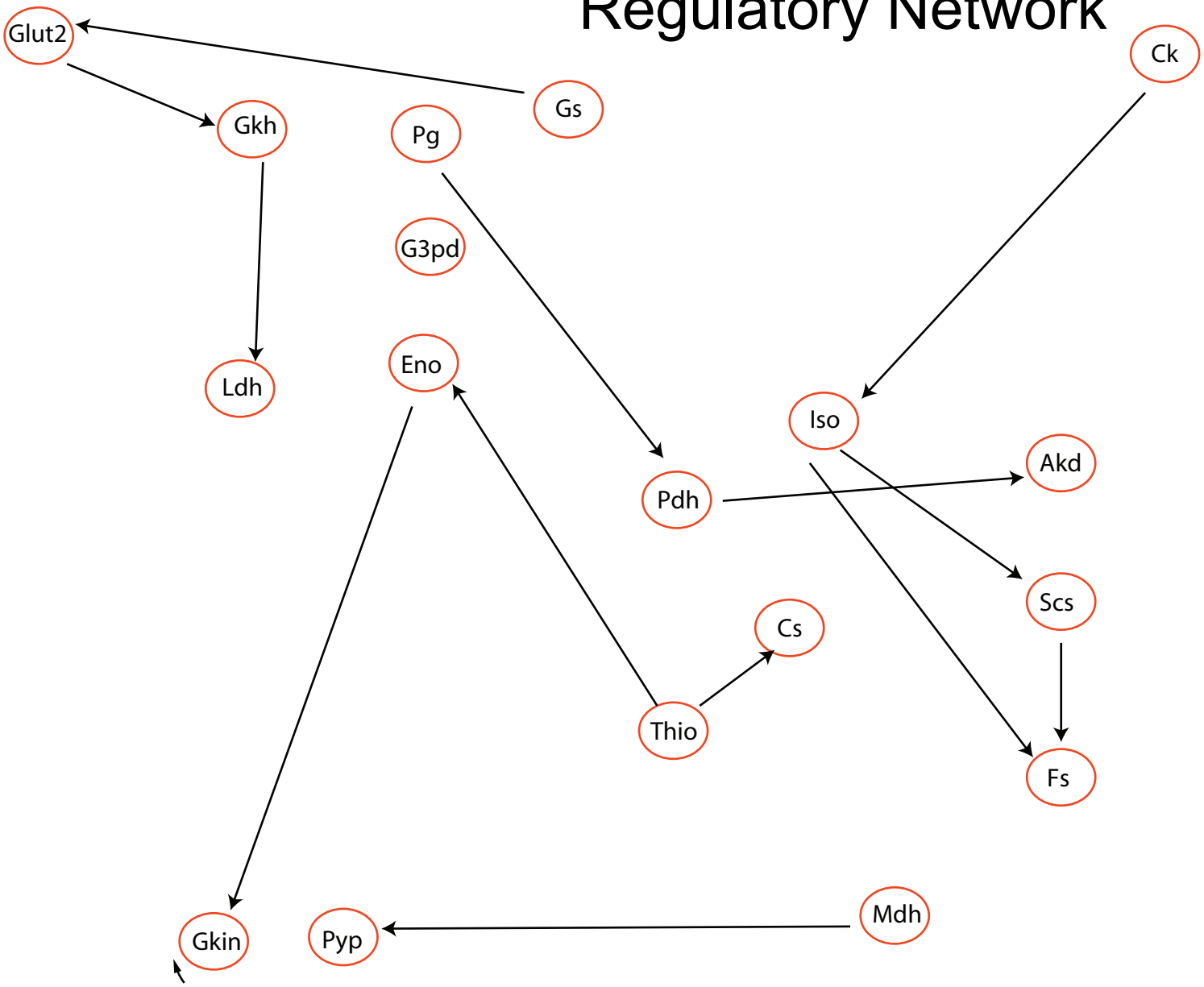
At steady state → Linear System


Cytosol
Mitochondria
Glycolytic Subdomain
Blood
Blood
GLU
G6P
GLY
GAP
BPG
LAC
PYR
PCr
Cr
ATP
ADP
bADP
FFA
FAC
GLR
TG
CIT
αKG
SCA
SUC
MAL
OAA
ACoA
CO_2
O_2
H_2O
P_i
CoA
NAD
NADH-H+
ADP+H+
ADP+2Pi
ADP+Pi
NAD+
NAD+Pi
NADH-H-
NAD-H+
H+
Pi
P_i+ADP

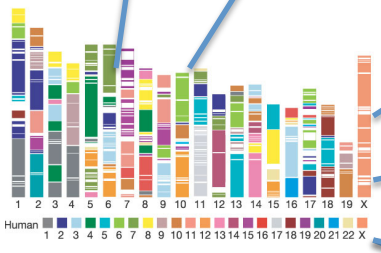

Cytosol
Mitochondria
Glycolytic Subdomain
Blood
Blood
Glut2
GLU
G6P
GLY
Gkh
Pg
Gs
GAP
G3pd
BPG
Eno
Ldh
LAC
PYR
ATP
ADP+H+
ADP+2Pi
Pi
ADP
NAD+
NADH-H+
NAD+Pi
NADH-H-
Ck
PCr
Cr
ATP
ADP
PYR
NAD
CoA
Pdh
CO2
ACoA
Cs
CIT
Iso
αKG
Akd
SCA
Scs
SUC
Fs
NAD-H+
OAA
Mdh
MAL
H+
bADP
Thio
FFA
FAC
Gkin
Pyp
GLR
TG
Pi+ADP
ADP+Pi
O2
H20
Pi
CoA


Glut2
Ck
Gkh
Gs
Pg
G3pd
Eno
Ldh
Iso
Akd
Pdh
Scs
Cs
Thio
Fs
Mdh
Gkin
Pyp

Regulatory Network

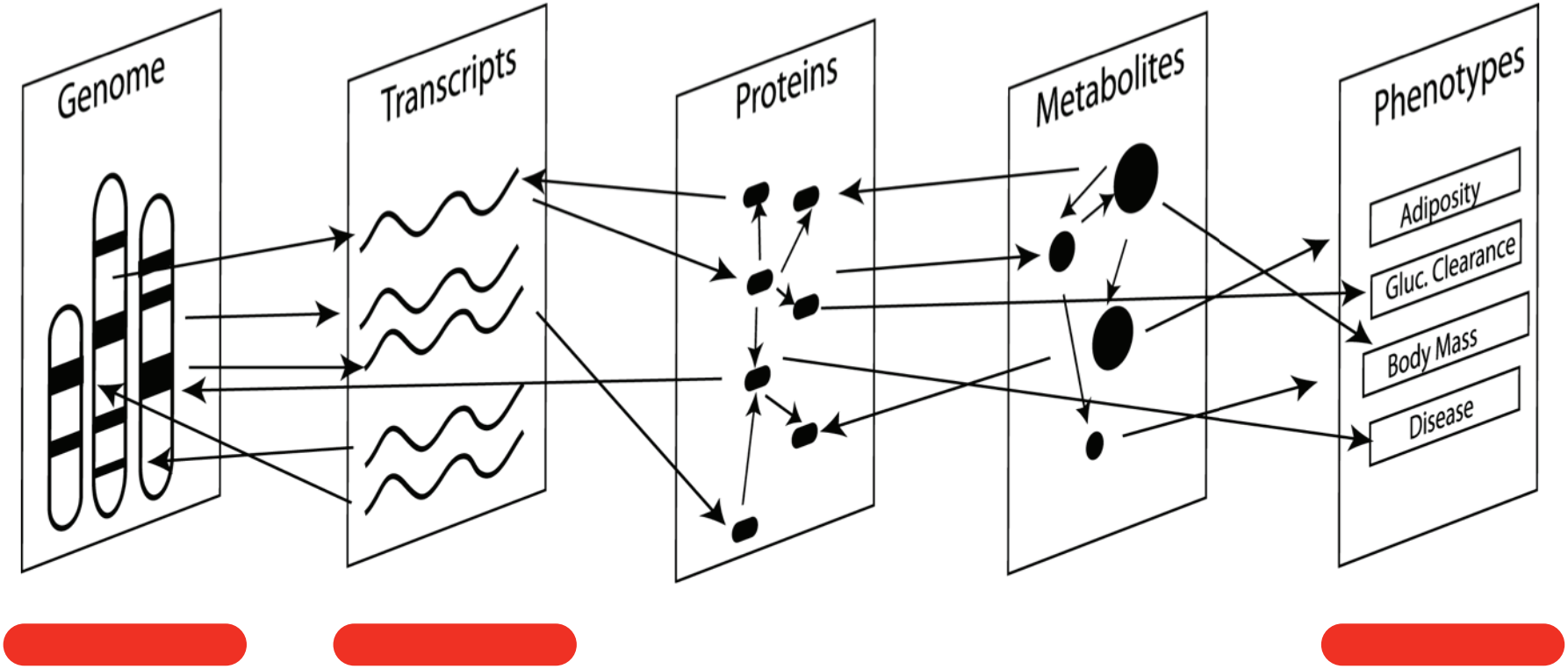


Regulatory Network



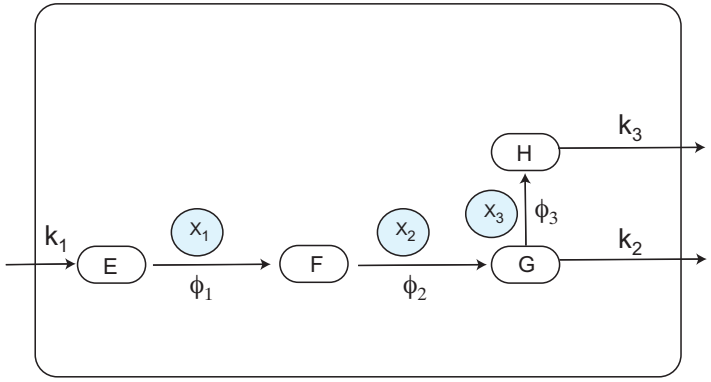
Glut2
Gs
Ck
Gkh
Pg
G3pd
Ldh
Eno
Iso
Akd
Pdh
Scs
Cs
Thio
Fs
Mdh
Gkin
Pyp

II. Causal Graphical Models

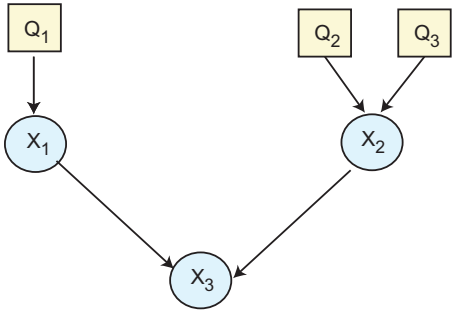


Merging Model Systems

Metabolic Model



Causal Graphical Model



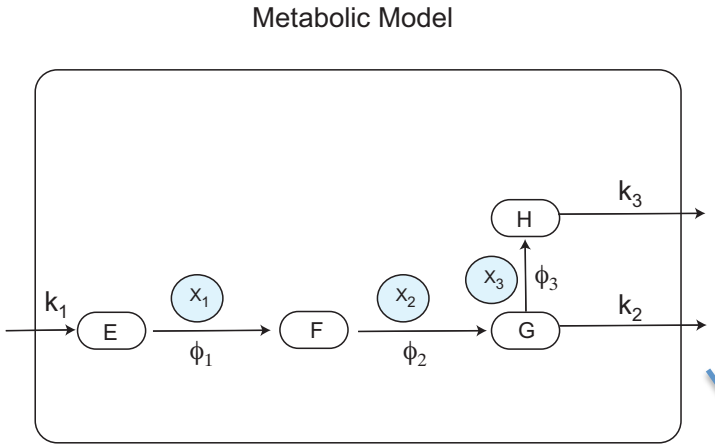
- Genotypes at QTL

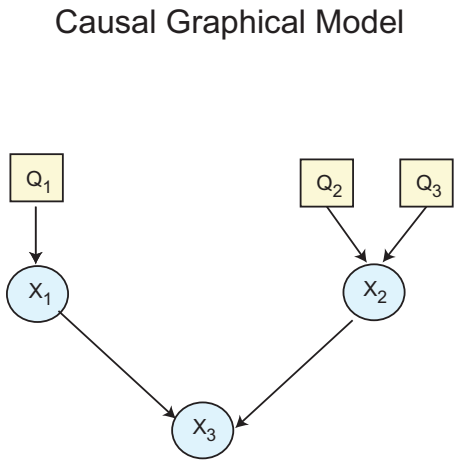
- Genes

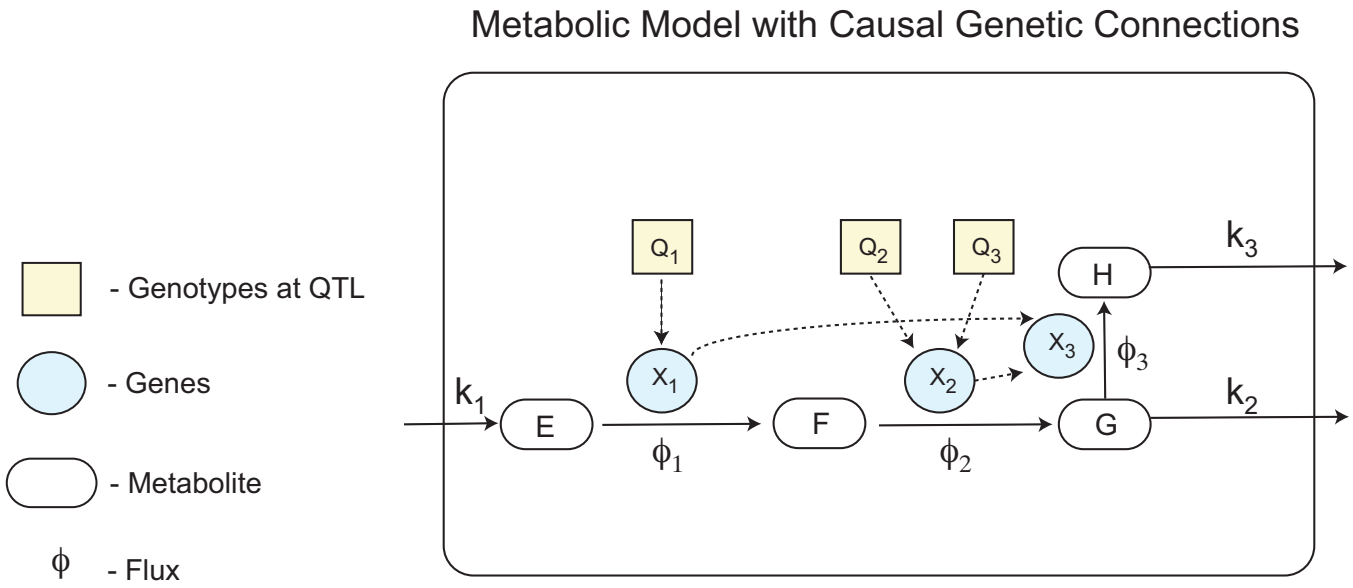
- Metabolite

ϕ - Flux

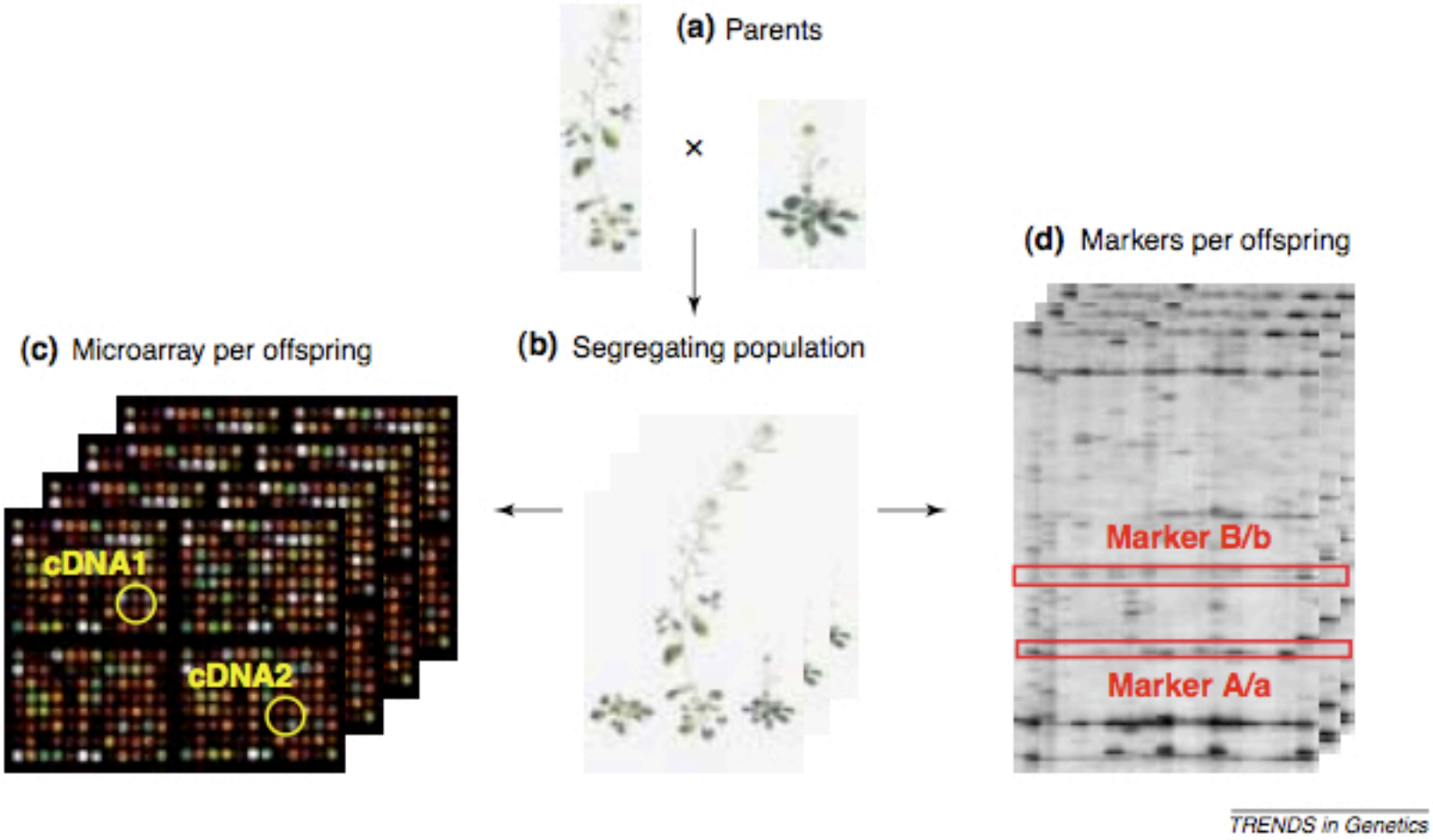
Merging Model Systems







Genetical Genomics



JANSEN, R. C., and J. NAP, 2001 Genetical genomics: the added value from segregation. Trends in Genetics **17:** 388-391.

Preliminaries

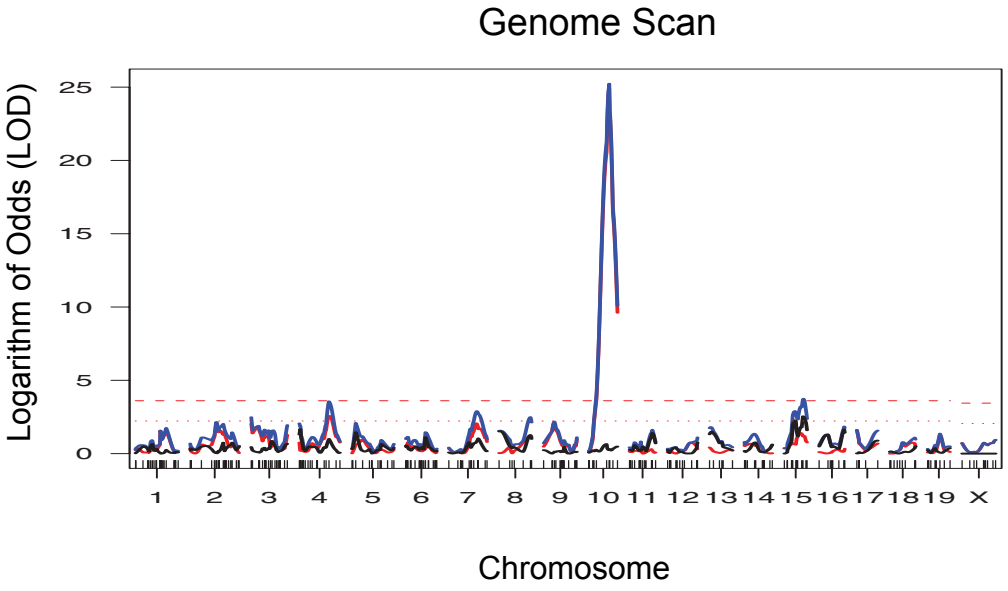
Quantitative Trait Locus (QTL):
A genomic region where allelic variation correlated with **trait** variation.

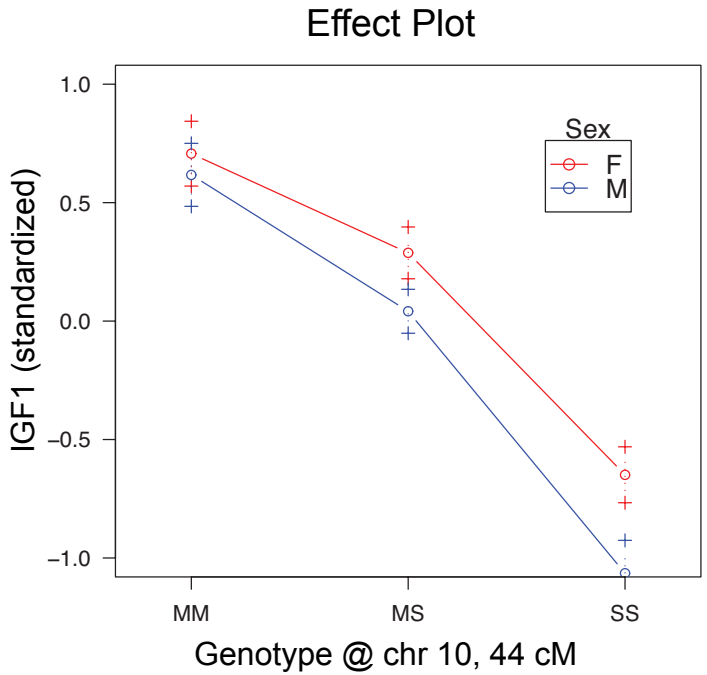
Trait:
Gene Expression (~40,000 transcript) and other clinical measurements.

Preliminaries

Quantitative Trait Locus (QTL):
A genomic region where allelic variation correlated with **trait** variation.

Trait:
Gene Expression (~40,000 transcript) and other clinical measurements.

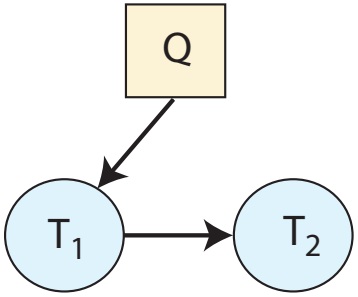




Leduc MS, Hageman RS, Meng Q, Verdugo RA, Tsaih SW, Churchill GA, Paigen B, Yuan R, Genomic analysis identifies loci regulating IGF1 level and longevity. Aging Cell 9(5): 823-836, (2010).

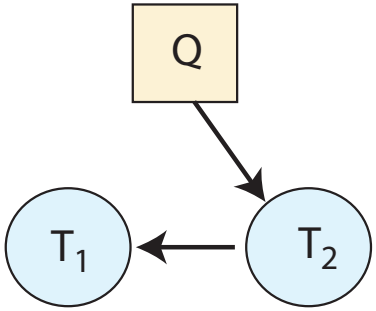
Local Approaches

Causal



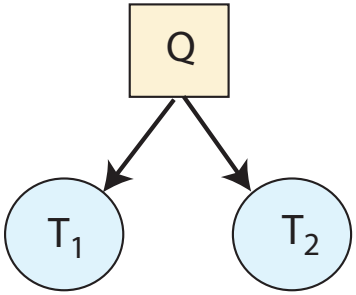
$$P(T_2 \mid T_1, Q) = P(T_2 \mid T_1)$$

Reactive



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

Independent



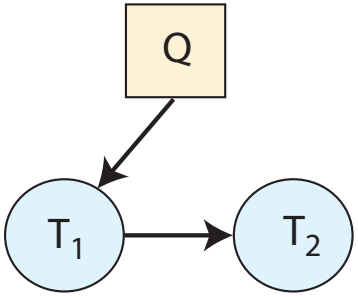
$$P(T_1 \mid T_2, Q) = P(T_1 \mid Q)$$

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

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nature 37: 710-717.

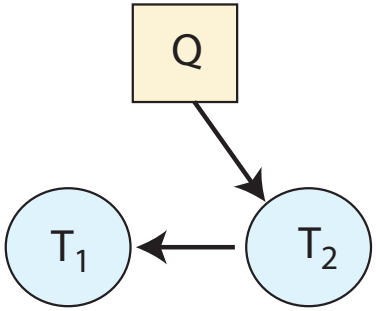
Local Approaches

Causal



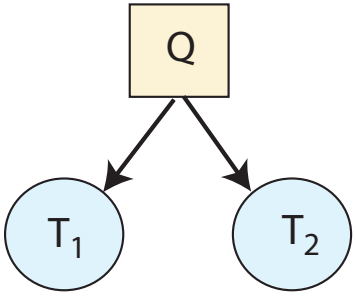
$$P(T_2 \mid T_1, Q) = P(T_2 \mid T_1)$$

Reactive



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

Independent



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

Limitations (the trouble with triplets):

- Identifies primary and secondary regulators – misses hierarchy of interactions.

- Local models pinned together = ‘hairball of traits’ -> over-fitting.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nature 37: 710-717.

Local Approaches

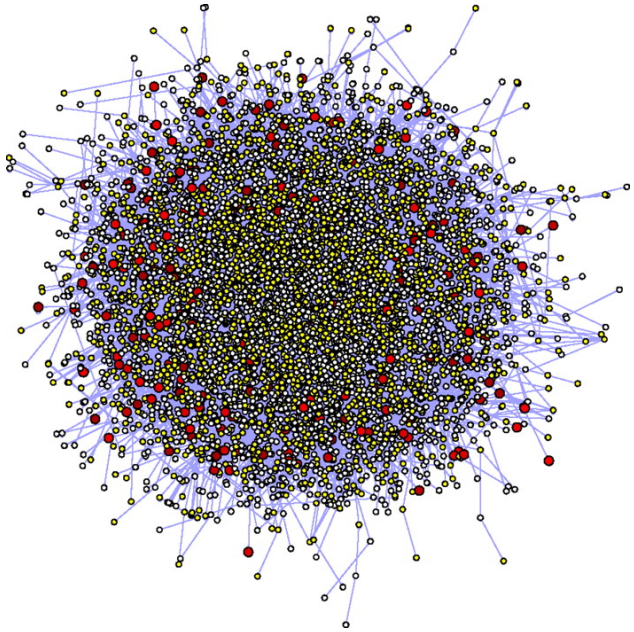


Image (left): http://www.pnas.org/content/104/51/20274/F1.large.jpg

Global Approaches

Rooted (loosely) in Probabilistic Graphical Models (PGMs):

- Homogeneous Conditional Gaussian Models
- Bayesian Networks
- Estimation of UDG, then directed.

Additional features (via priors):

- Penalty on graph density.
- Restrictions on the number of parent nodes (fan-in).

Structural learning: greedy or MCMC-based.

Bayesian Networks

The Data: phenotypes (X) and genotypes (Q):

$$D = \left\{ \underbrace{X_1, X_2, \ldots X_n,}_{phenotypes} \ldots, \underbrace{Q_1, Q_2, \ldots Q_m}_{genotypes} \right\}.$$

The Assumption: The graph (G) is a Directed Acyclic Graph (DAG).

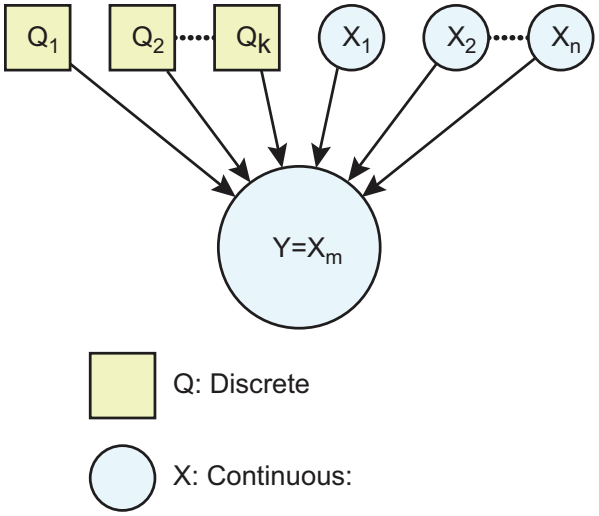
$$P\left(D_1, D_2, \ldots, D_{n+m}\right) = \prod_{i=1}^{k} P\left(D_i \mid \pi_G(D_i)\right)$$

The Posterior Probability:

$$P(G \mid D) \; \alpha \prod_{i=1}^{k} \underbrace{P(D \mid G)}_{likelihood} \; \underbrace{P(G)}_{structural\ prior} \quad ,$$

$$\text{where} \;\; P(D \mid G) \alpha \int P(D \mid \theta, G) P(\theta \mid G) d\theta.$$

Local Families



The model:

A continuous child $y = X_m$ with parents $\pi_G(y) = \{Q_1,\ldots,Q_k,X_1,\ldots,X_n\}$ is modeled as:

$$Y = \beta_0 + \beta_1 Q_{A,i} + \beta_2 Q_{B,i} + \beta_3 Q_{H,i} + \ldots + \beta_{s-2} Q_{A,k} + \beta_{s-1} Q_{B,k} + \beta_s Q_{H,k} + \beta_{s+1} X_1 + \ldots + \beta_t X_n + \varepsilon,$$

where $\beta_i \sim N\left(\mu_i, \sigma_i^2\right)$

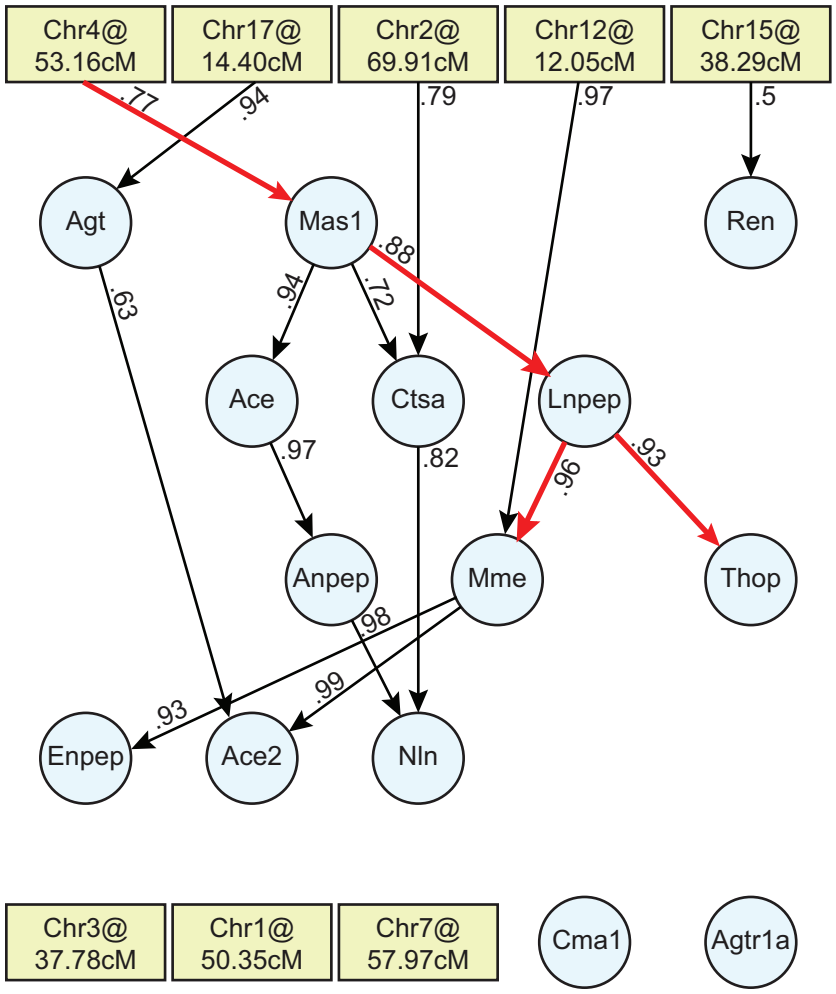
Assumption:

Each child can have at most *k* parents.

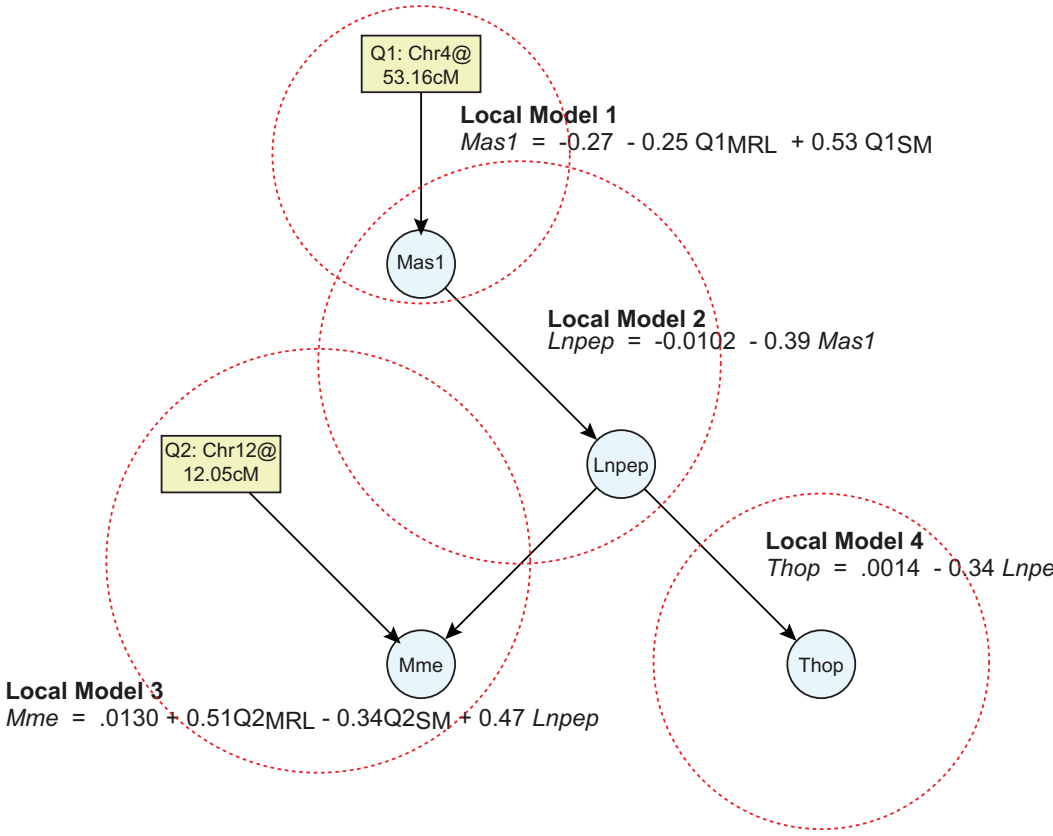
GELMAN, A., and J. HILL, 2007 *Data Analysis using Multilevel/Hierarchical Models*. Cambridge University Press.

Sample Output

Marginal Summary

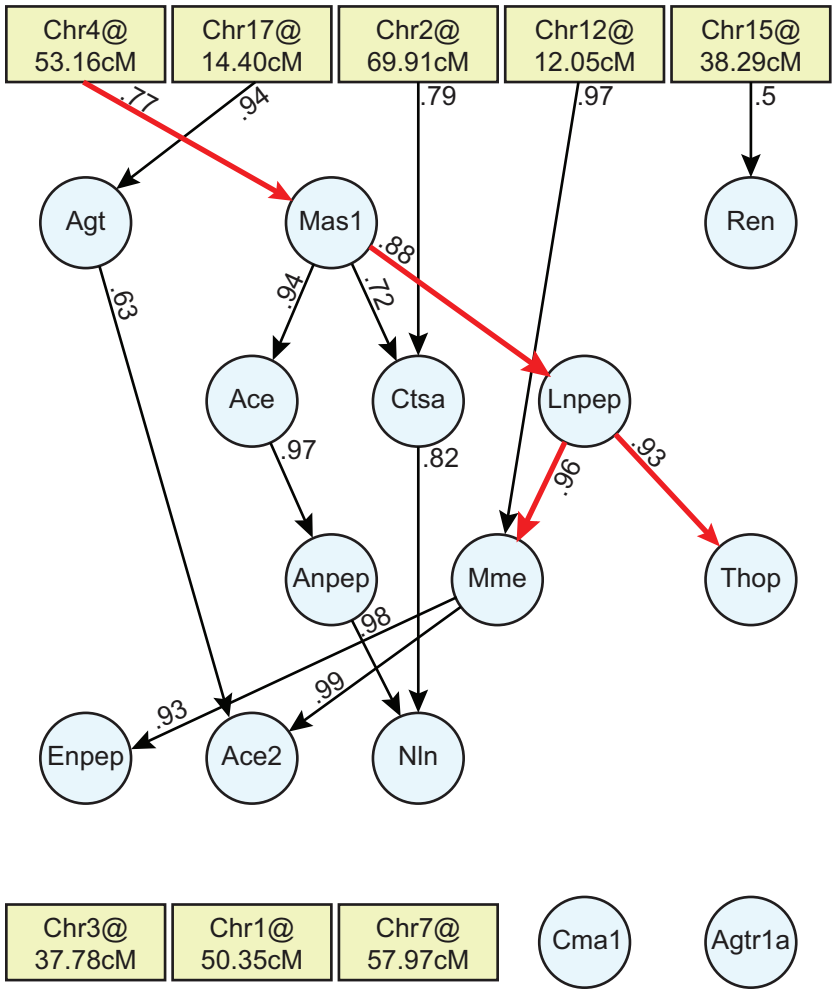


Local Models

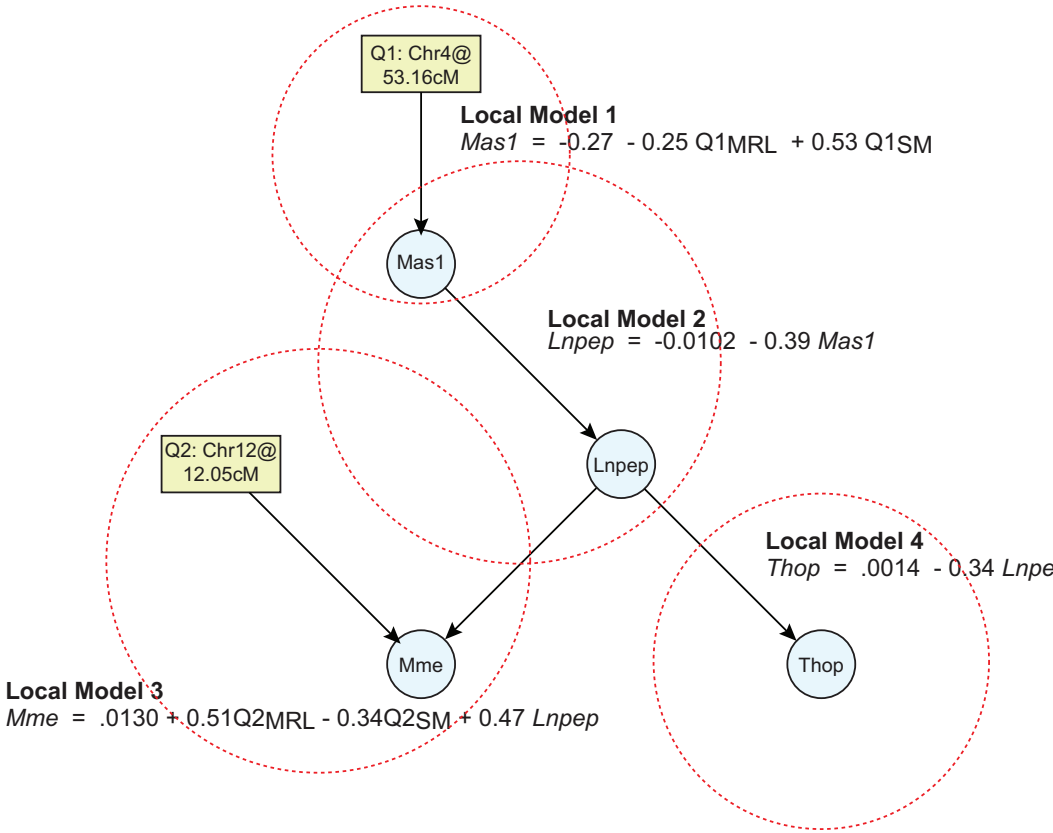


Sample Output

Marginal Summary



Local Models



Now What?

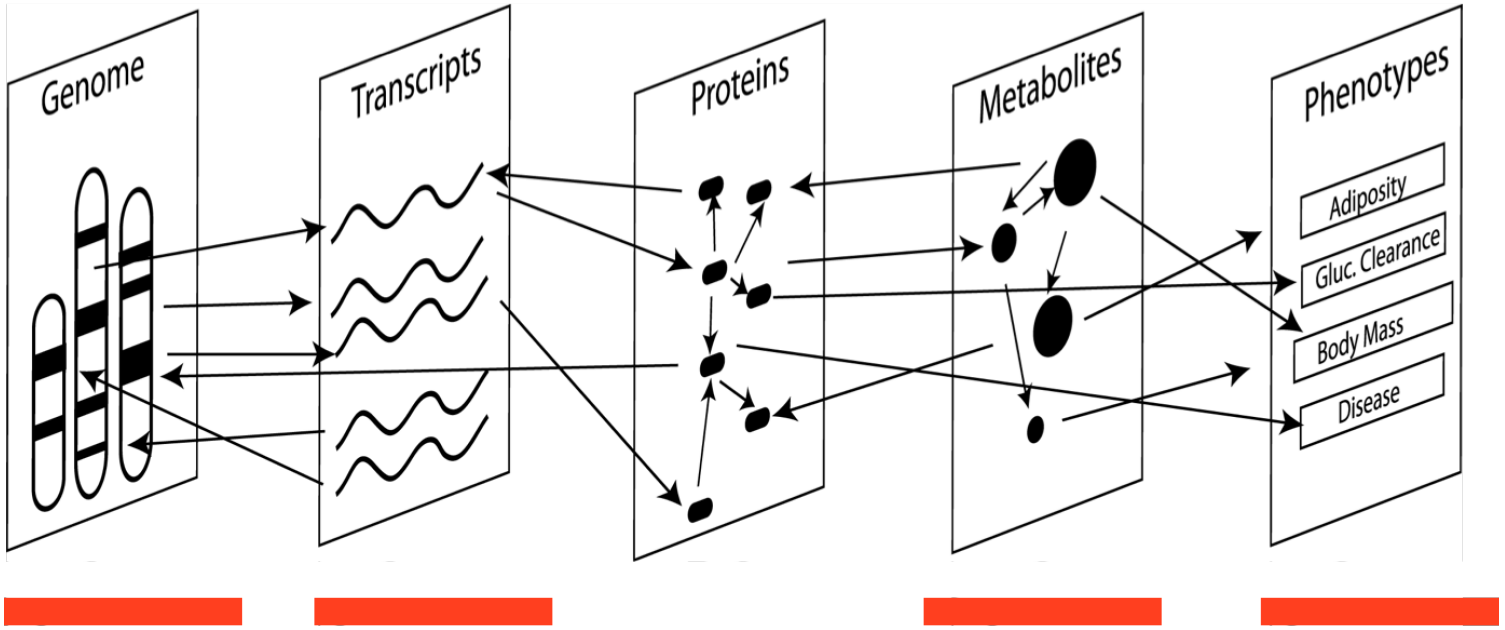
Now What?

- 12 + approaches to genotype-phenotype network inference.

- Network structure has been the “endpoint”.

- Limitations to “model interpretation”.
  Can visually detect “direct” and “indirect” relationships.
  Can attempt to quantify “strength” of the relationship.

III. Belief Propagation in Genotype-Phenotype Networks



- 12 + approaches to genotype-phenotype network inference.
- Network structure has been the "endpoint".
- Limitations to "model interpretation".

  Can visually detect "direct" and "indirect" relationships.

  Can attempt to quantify "strength" of the relationship.

Belief Propagation: Motivation

Question: Suppose we have “new information” about the system (e.g., a knock-out of a gene, a genotype of an individual, or a phenotype value). Can we understand the system-wide response to this “new information”?

Changing our way of thinking:

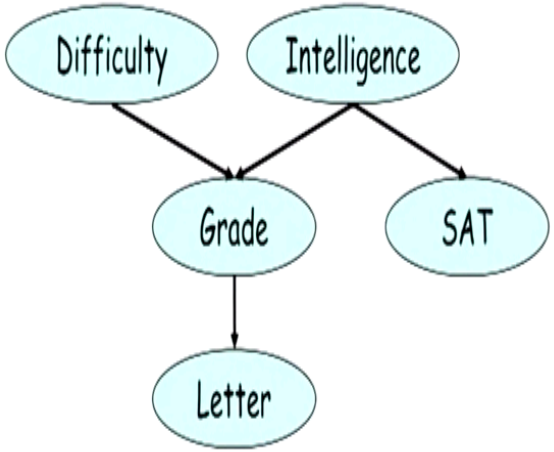
Before: knock out gene A -> everything downstream is effected (unclear how exactly).

Now: : knock out gene A -> all nodes that are d-connected to A will be effected (Causal reasoning, Evidential Reasoning, Inter-causal reasoning).

Belief Propagation: Background

When can X influence Y?

- X -> Y - yes, straight downward path
- X <- Y – yes, evidential reasoning
- X -> W -> Y – yes
- X <- W <- Y - yes
- X <- W -> Y - yes
- X -> W <- Y - no V-structure/collider model



Belief Propagation: Background

A trail $X_1 - X_2 - \ldots - X_k$ is **active** if it has no v-structures:

$$X_{i-1} \rightarrow X_i \leftarrow X_{i+1}.$$

General edge type up/down

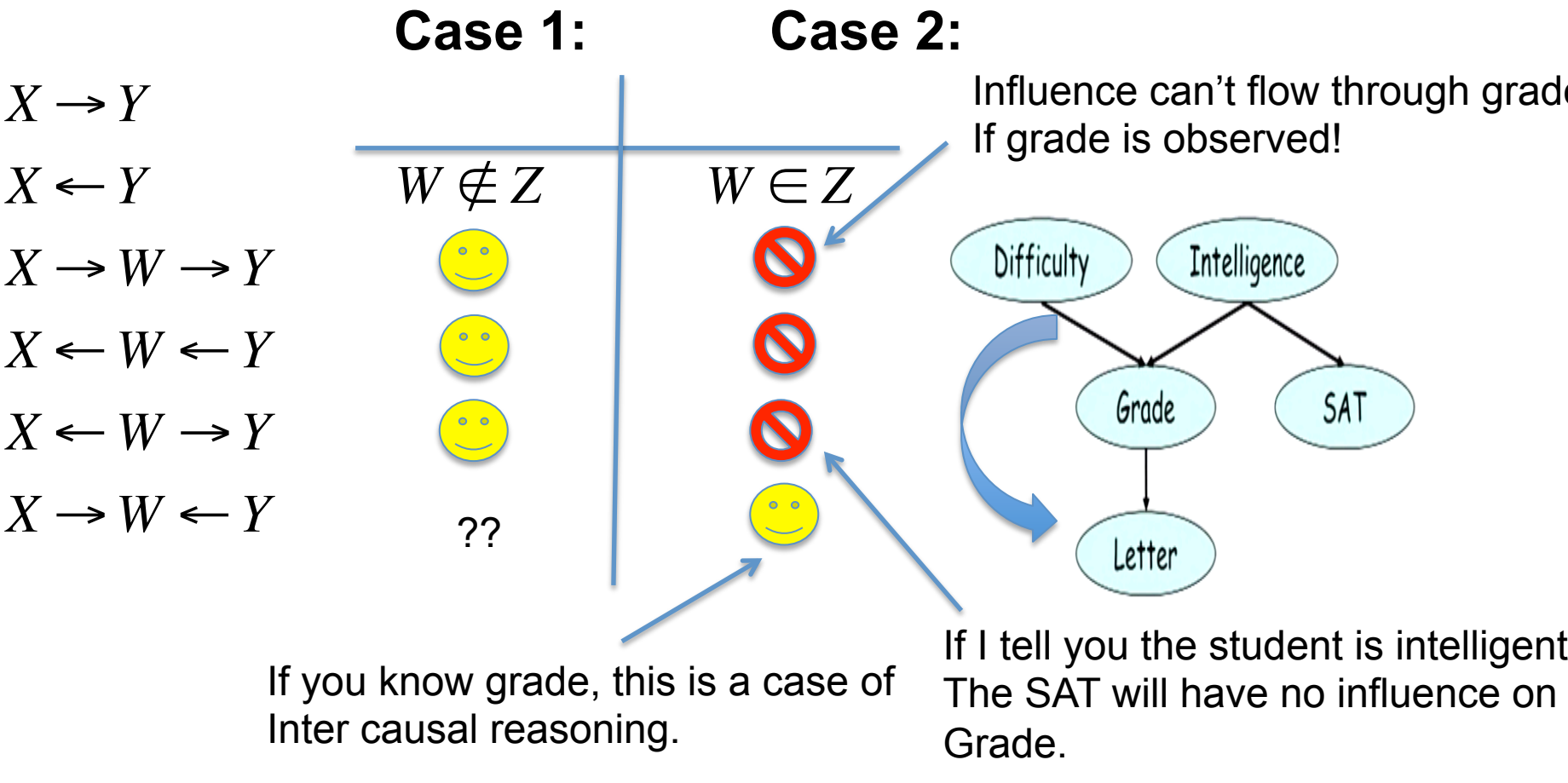
A block in the trail

Therefore, information can flow freely though the network, in the "active" sense, unless a v-structure arises.

**Lets think about additional evidence*

Belief Propagation: Background

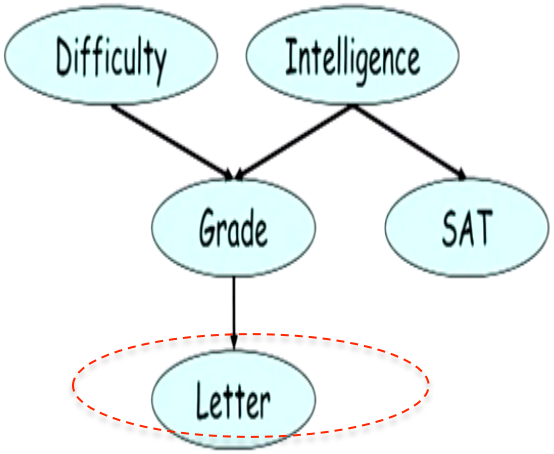
When can X influence Y given evidence about Z?:



Belief Propagation: Background

When can X influence Y given evidence about Z:

	Case 1: $W \notin Z$	Case 2: $W \in Z$
$X \rightarrow Y$		
$X \leftarrow Y$		
$X \rightarrow W \rightarrow Y$		
$X \leftarrow W \leftarrow Y$		
$X \leftarrow W \rightarrow Y$		
$X \rightarrow W \leftarrow Y$		



If you know don't observe grade directly, but you observe letter?

- if W and all of its descendants are not observed.
- if W or one of its descendants is observed.

Belief Propagation: Background

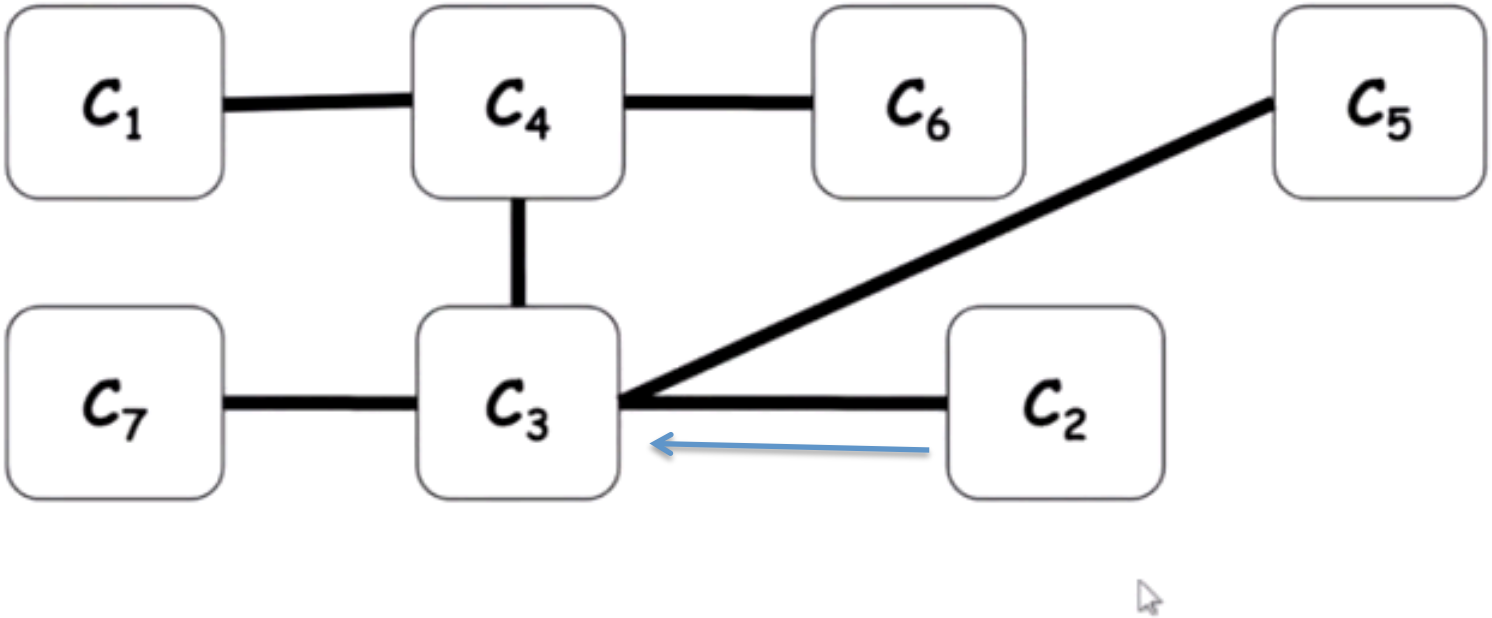
Definition: X and Y are d-separated in G given Z if there is no active trail in G between X and Y given Z

Notation:

$$\text{d-sep}_G(X, Y \mid Z)$$

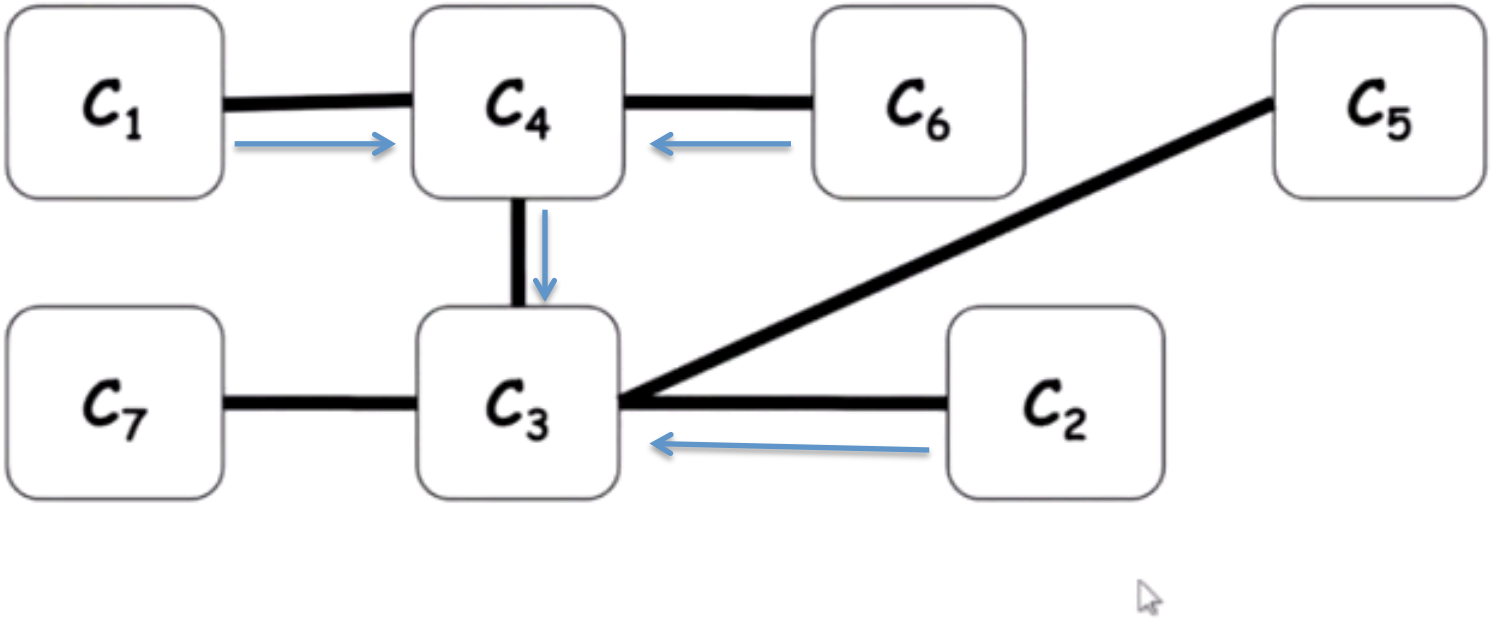
Belief Propagation: Background

- Message Passing order: we can start with any leaf.



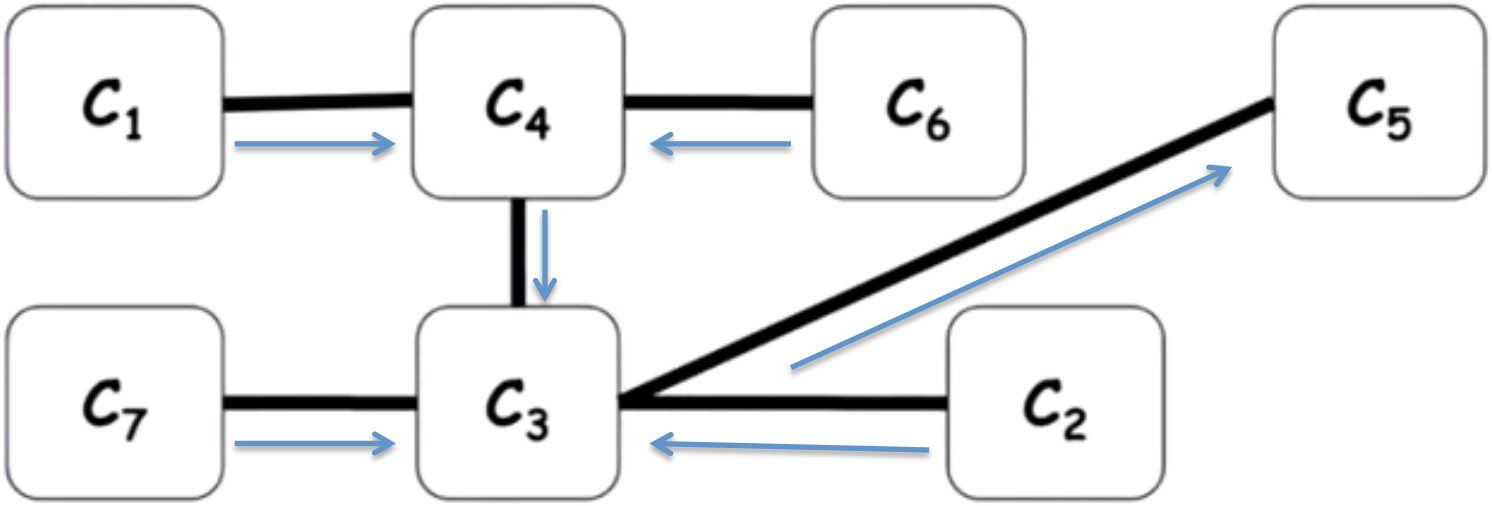
Belief Propagation: Background

- Message Passing order: we can start with any leaf.



Belief Propagation: Background

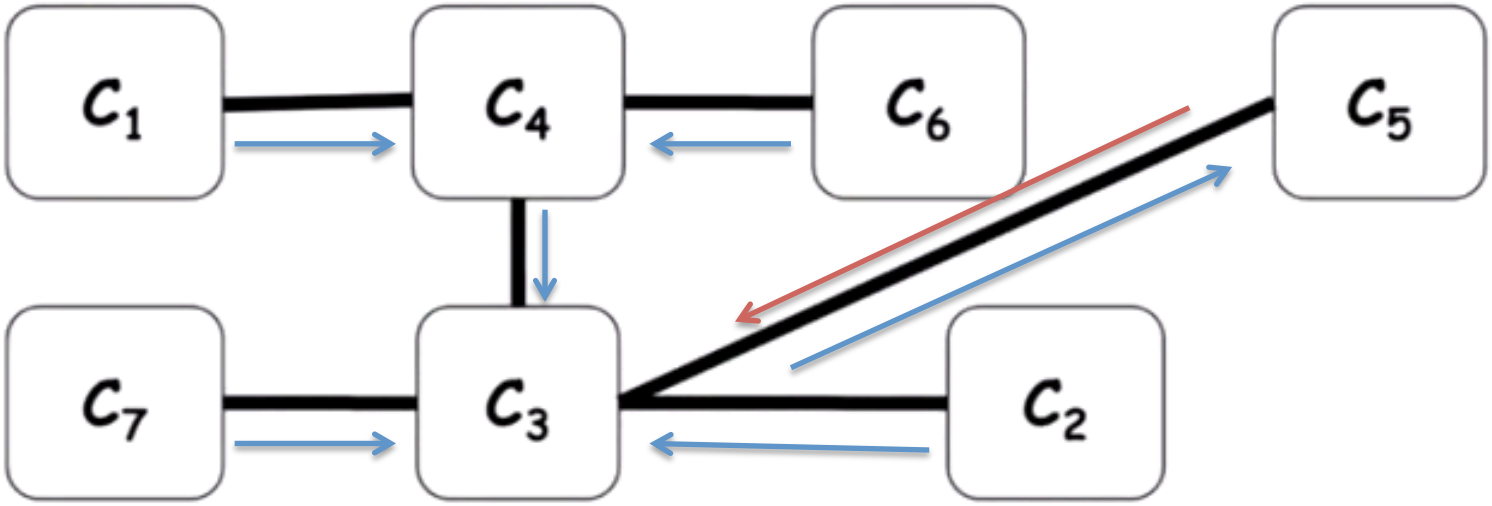
- Message Passing order: we can start with any leaf.



Everyone has received a message!

Belief Propagation: Background

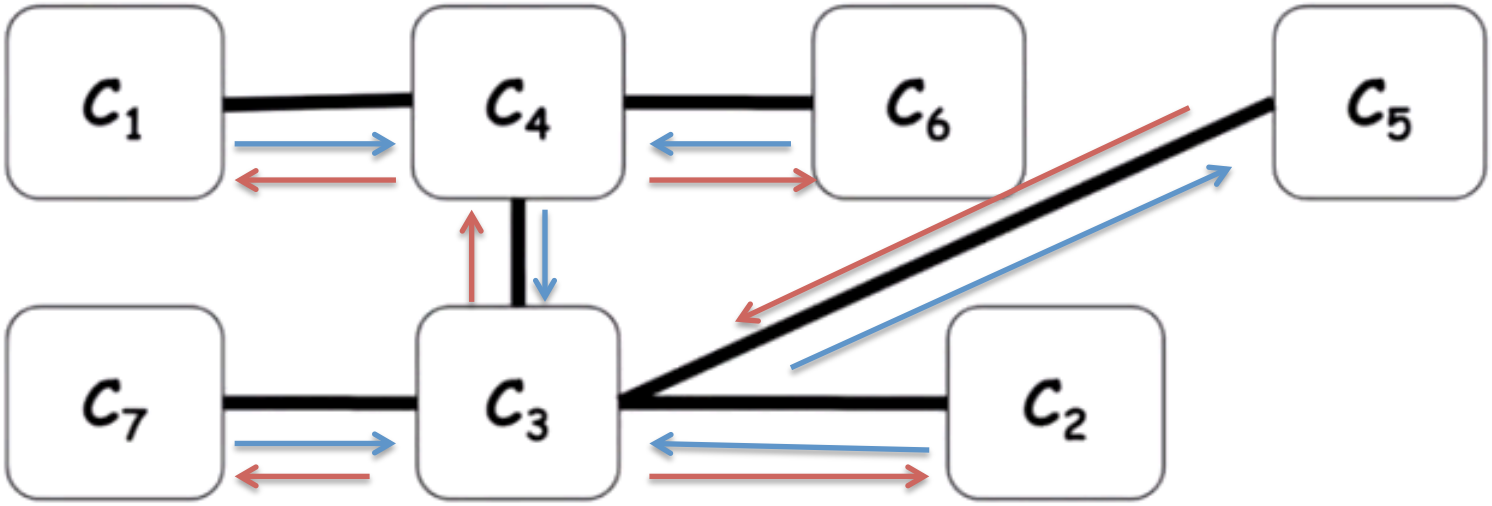
- Message Passing order: we can start with any leaf.



Everyone has received a message!

Belief Propagation: Background

- Message Passing order: we can start with any leaf.

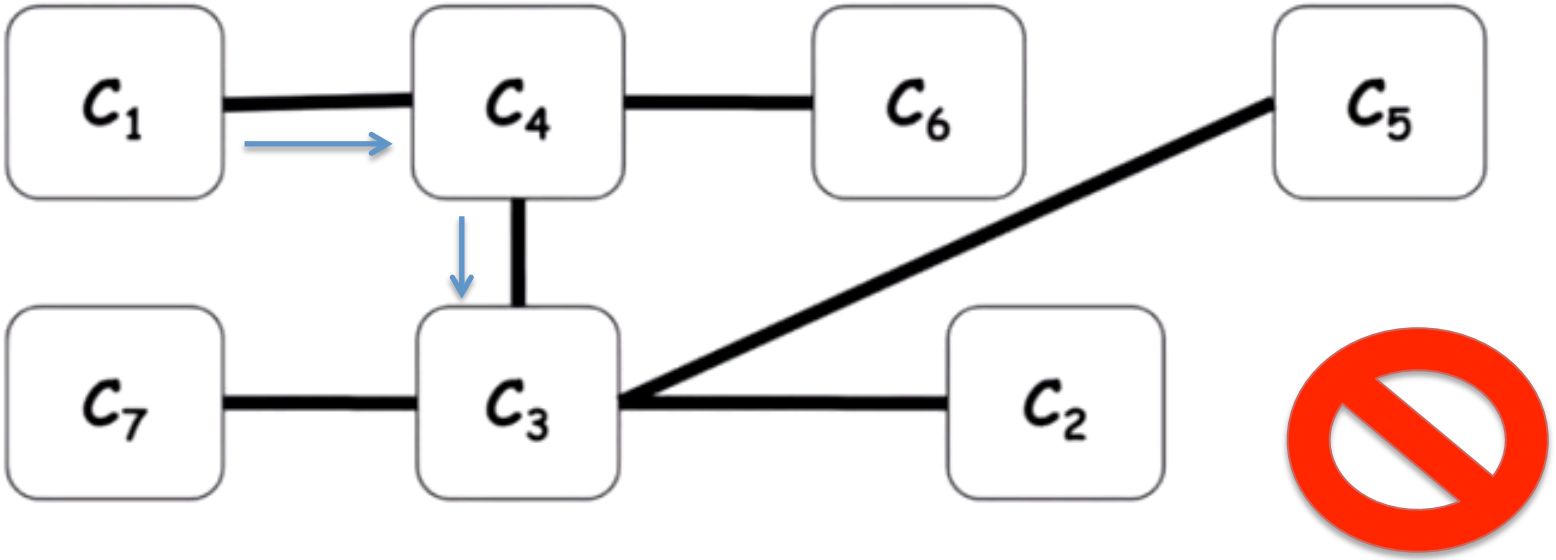


Everyone has received a message!

Everyone has passed a message!

Belief Propagation: Background

- Message Passing order: we can start with any leaf.



Illegal! C4 has to wait to gather
All of its information before talking.

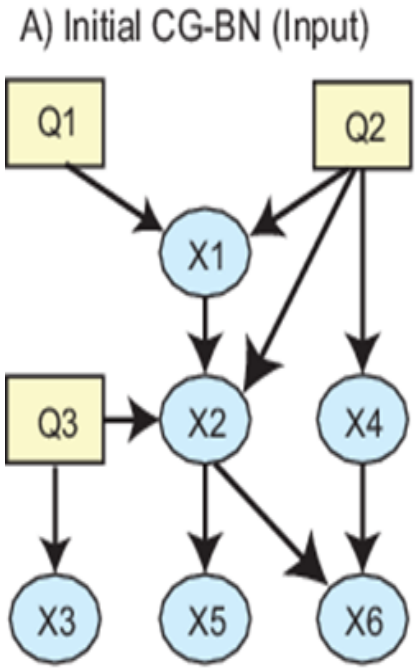
Belief Propagation: Background

Introducing new evidence Z=z, and querying X.

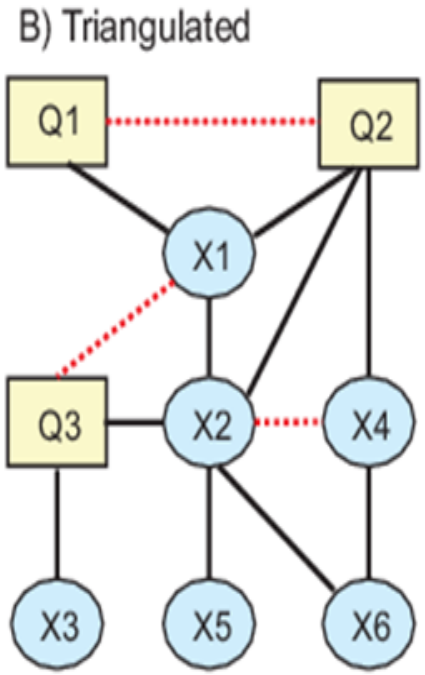
- **Case 1**: If X appears in clique with Z.
  Multiply clique that contains X and Z with indicator function 1(Z=z). (Reduce evidence).
  To get posterior, sum out irrelevant variables and renormalize.

- **Case 2**: If X does not appears in clique with Z.
  Multiply clique that contains X and Z with indicator function 1(Z=z). (Reduce evidence).
  Change the messages…. And pass on!

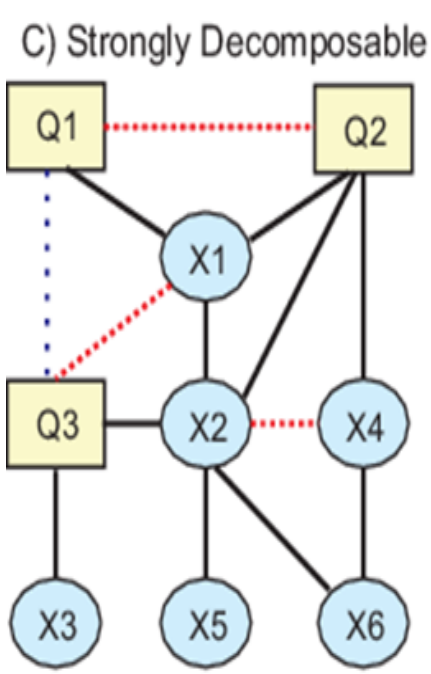
Belief Propagation in Genotype-Phenotype Networks



Start with a known network structure

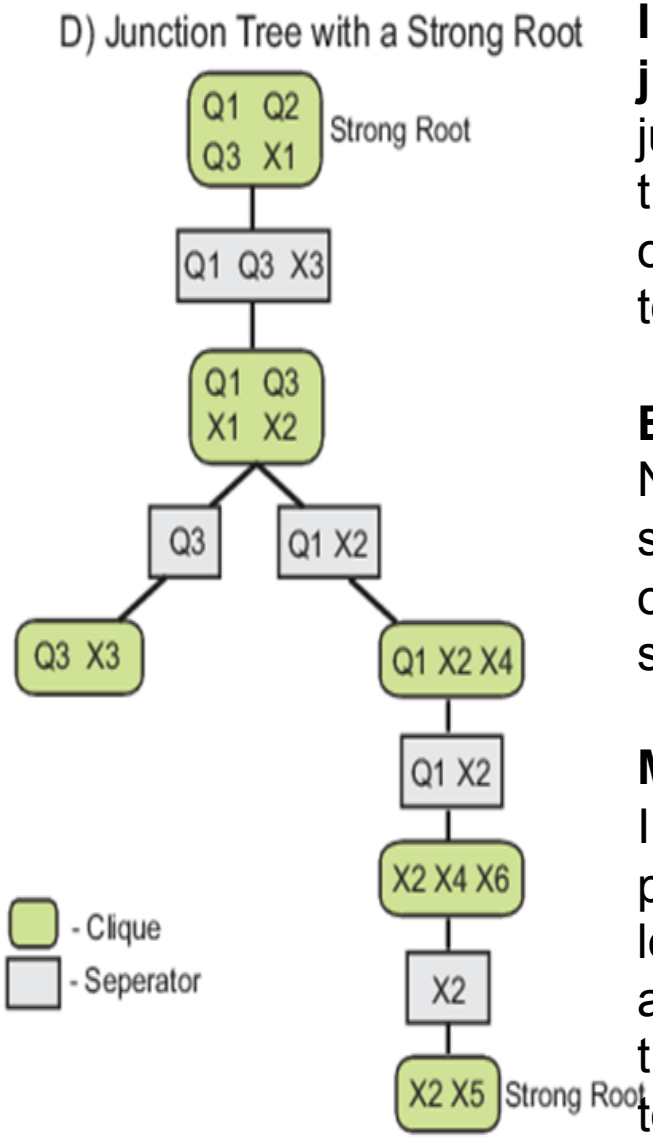


Marry the parents and drop directionality



No two discrete nodes be connected by a path that passes only through continuous nodes

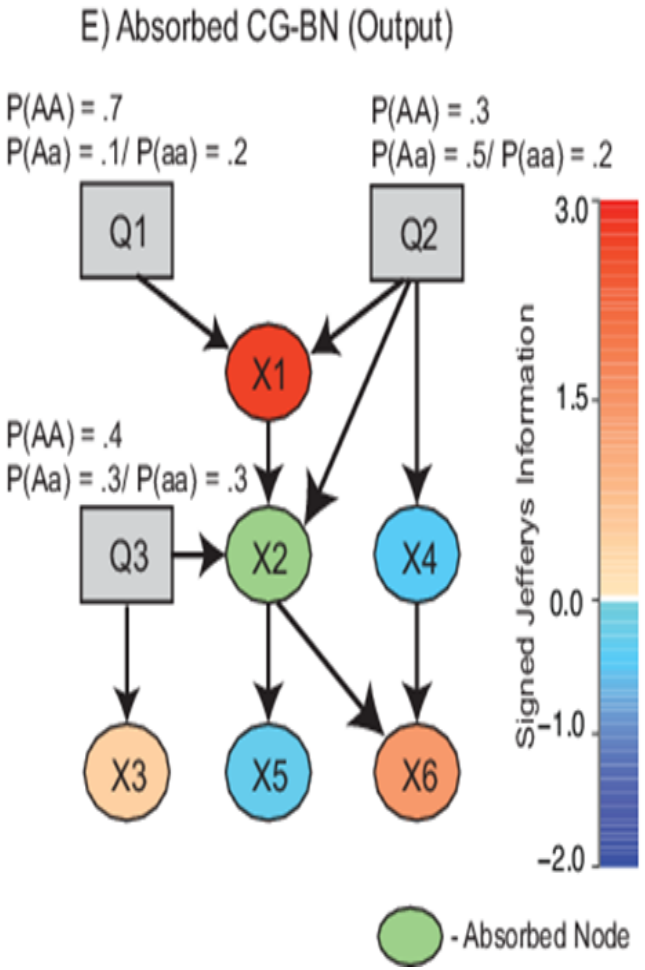
Belief Propagation in Genotype-Phenotype Networks



Initialization of the junction tree: The junction tree is initialized through the assignment of each node, X_i and Q_i, to a universe, V

Evidence Absorption: New evidence entered by setting phenotype $X_i = x_i^*$ or setting a genotype state $Q_i = g^*$

Message Passing: Information is rest propagated from the leaves to the strong root, and then distributed from the strong root back out to the leaves of the tree



Belief Propagation in Genotype-Phenotype Networks

- Predicting the how the network changes under new lines of evidence(s).

- **Initial State**: network with no absorbed evidence.

  Absorbed State: network after absorbed evidence is propagated.

Distance between initial and absorbed states: measured via signed Jeffery's Information (symmetric version of Kullbeck Lieber distance):

Jeffrey's information, which is computed for all continuous unabsorbed nodes in the network, is given as:

$$J\left(X_i^0, X_i^{\text{abs}}\right) = I^{\text{KL}}\left(X_i^0, X_i^{\text{abs}}\right) + I^{\text{KL}}\left(X_i^{\text{abs}}, X_i^0\right)$$

where

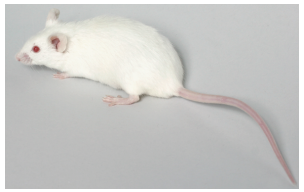
$$I^{\text{KL}}\left(X_i^0, X_i^{\text{abs}}\right) = \frac{1}{2}\left\{\frac{(\mu_0 - \mu_{\text{abs}})^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma_{\text{abs}}^2} - \log\left(\frac{\sigma_0^2}{\sigma_{\text{abs}}^2}\right) - 1\right\}.$$

For ease of interpretation, the signed Jeffrey's information:

$$\text{sign}(\mu_0 - \mu_{\text{abs}}) \cdot \text{J}\left(\text{X}_\text{i}^0, \text{X}_\text{i}^{\text{abs}}\right)$$

Application

MRL/MpJ



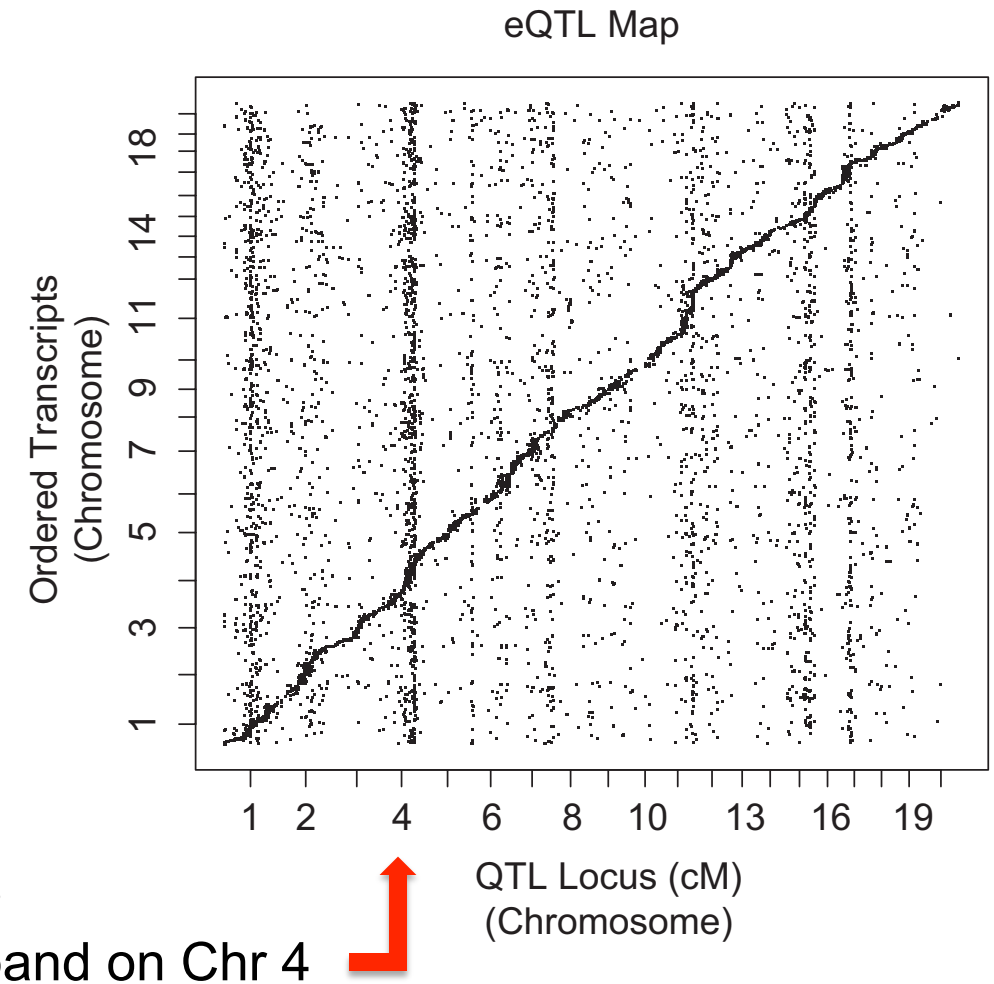
X

SM/J

Kidney Disease susceptibility

173 Male Kidneys

Measurements:
~ 35,000 gene expression traits
~ 15 clinical traits
~ genotypes at 256 SNP markers



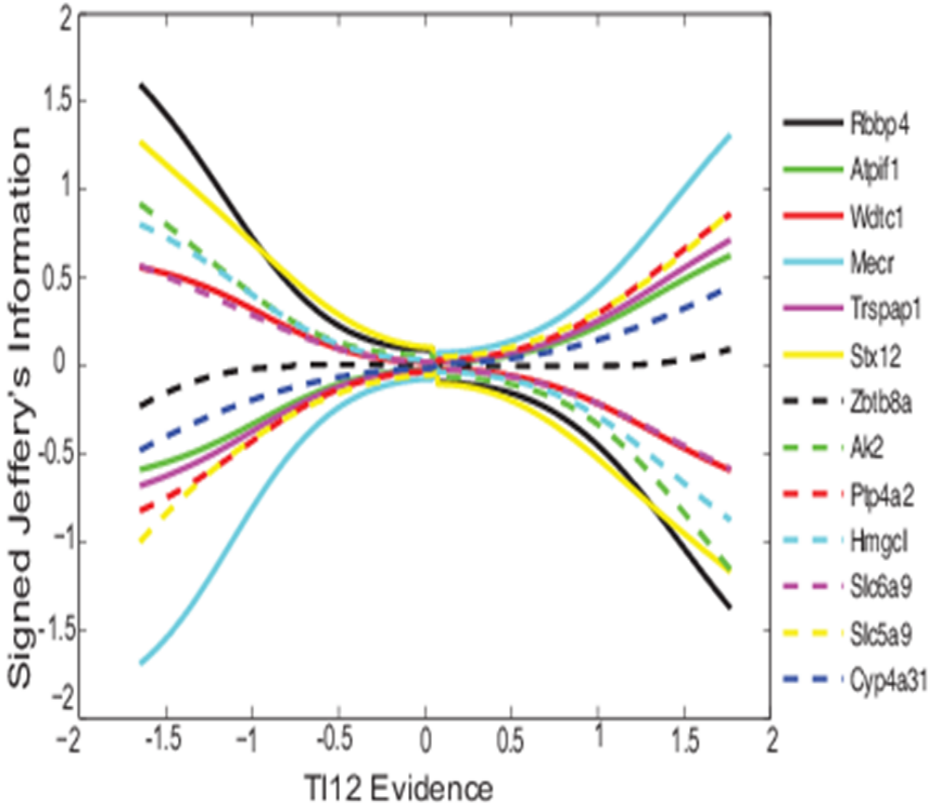
Trans-band on Chr 4
Enriched for Renin-Angiotensin System (RAS)

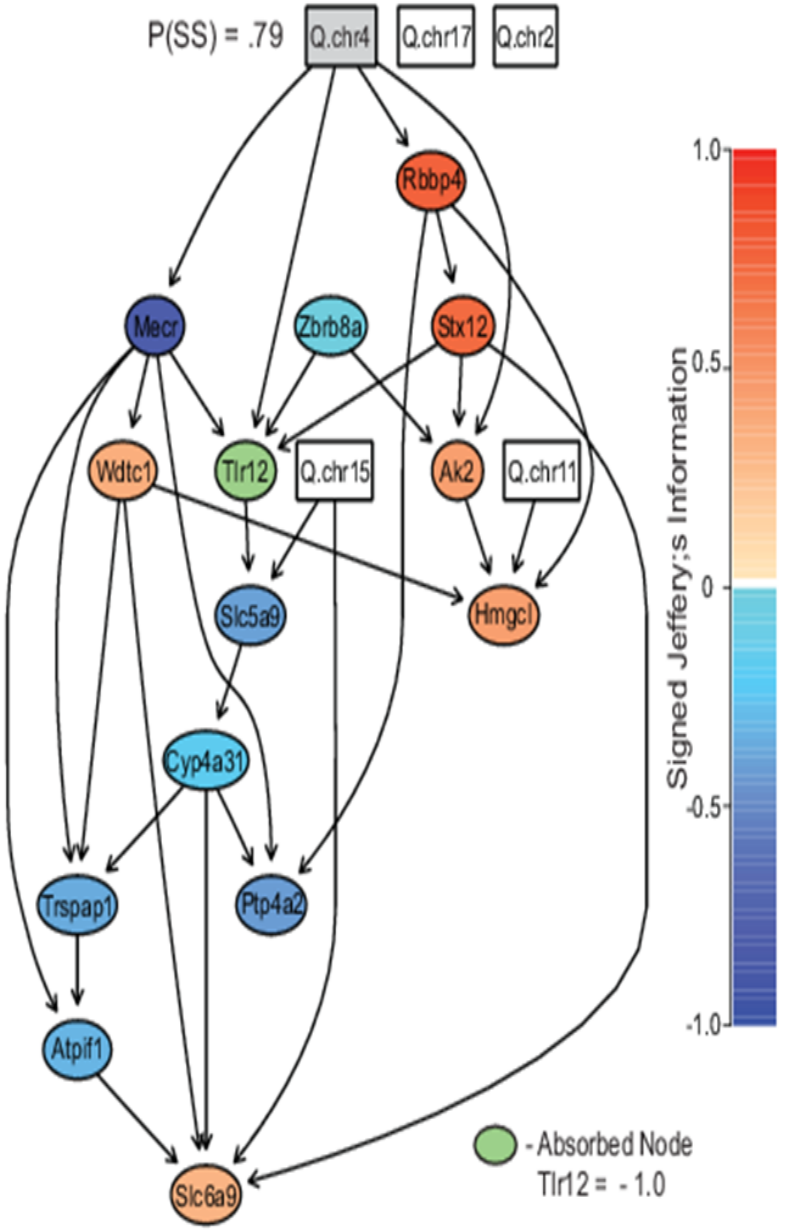
Hageman RS, Leduc MS, Caputo CR, Tsaih SW, Paigen B, Churchill GA, and Korstanje R. Uncovering Genes and Regulatory Pathways Related to Urinary Albumin Excretion in Mice. Journal of the American Society of Nephrology *22: 73-81 (2011).*

Application

- Mus musculus Kidney eQTL Data:
  - 173 males, F2 inter-cross between inbred MRL/MpJ and SM/J strains of mice.
- Pre-processing
  - Variable selection performed by filtering on significance and location of QTL, followed by a cross-validated elastic net procedure, with *Tlr12* as the response.
  - The 14 genes and their SNP markers corresponding to their QTL were included as variables for the graphical model.
- Structure Learning
  - PC-algorithm using *RHugin* package for the R programming language (α incremented from 0 to 0.1)
  - QTLnet method (Markov chain of 20000 iterations, burn-in rate 10%, model averaged network structure constructed from causal relationships with posterior probability of 0.5 or higher)

Results: A single line of evidence





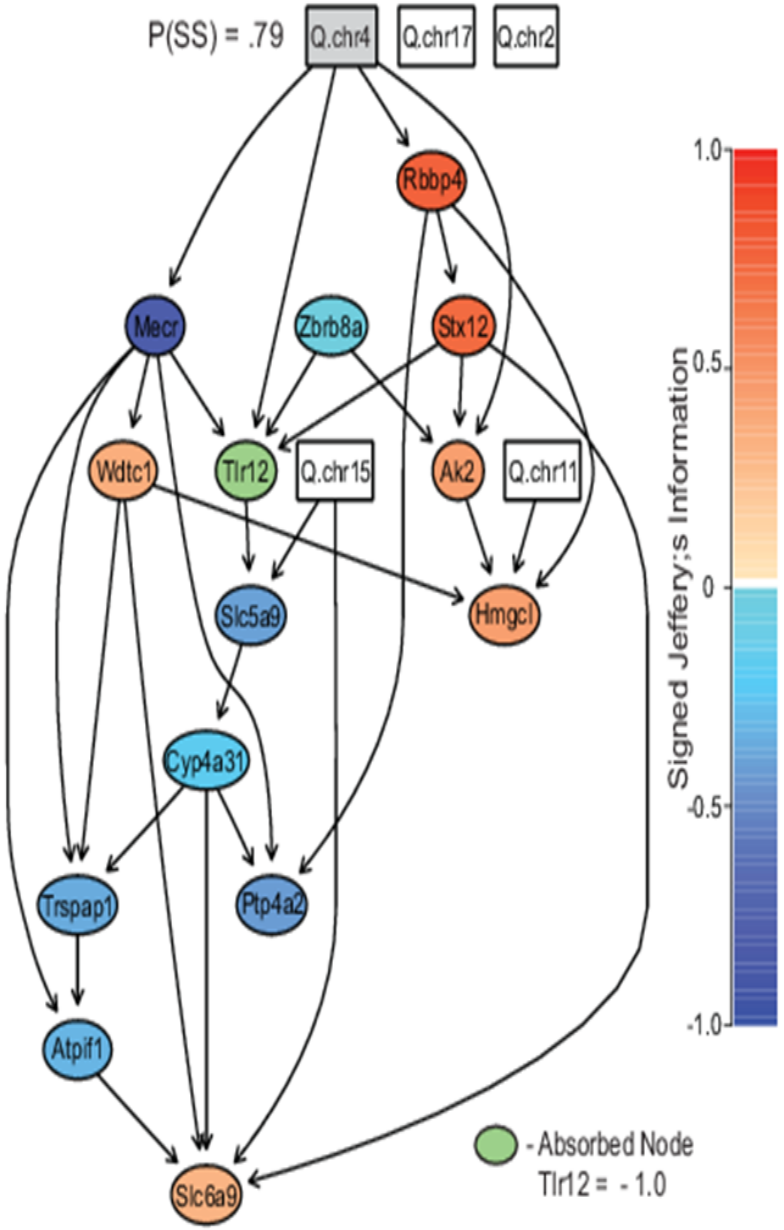
Results: A single line of evidence

Coordination and co-regulation are suggested in the direction of effect observed in the different regions of the pathway

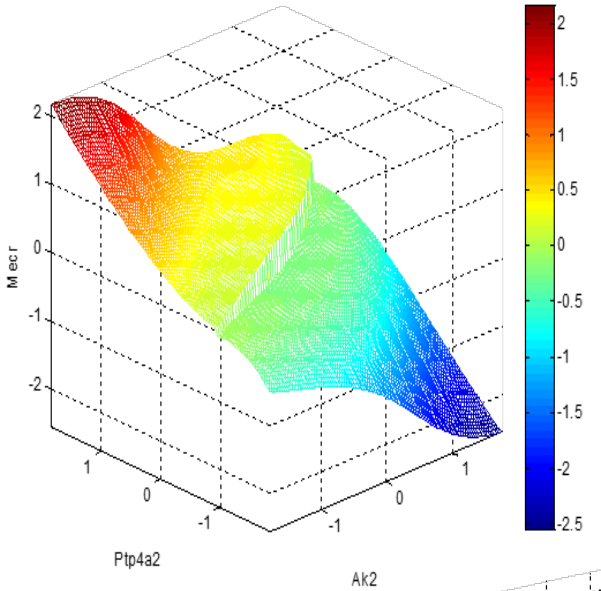
- **Activation** of {***Rbbp4, Stx12, Ak2, Hmgcl***} genes involved either in AMP/ADP/ATP metabolism or protein biosynthesis/transport
- **Repression** {***Mecr, Zbtb8a, Slc5a9, Cyp4a31, Ptp4a2, Trspap1, Atpif1***}

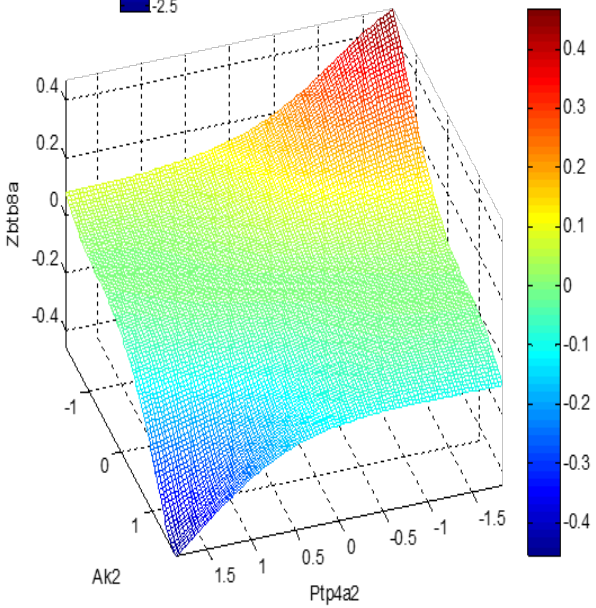
Absorbing evidence in *Tlr12* < 0 leads to:

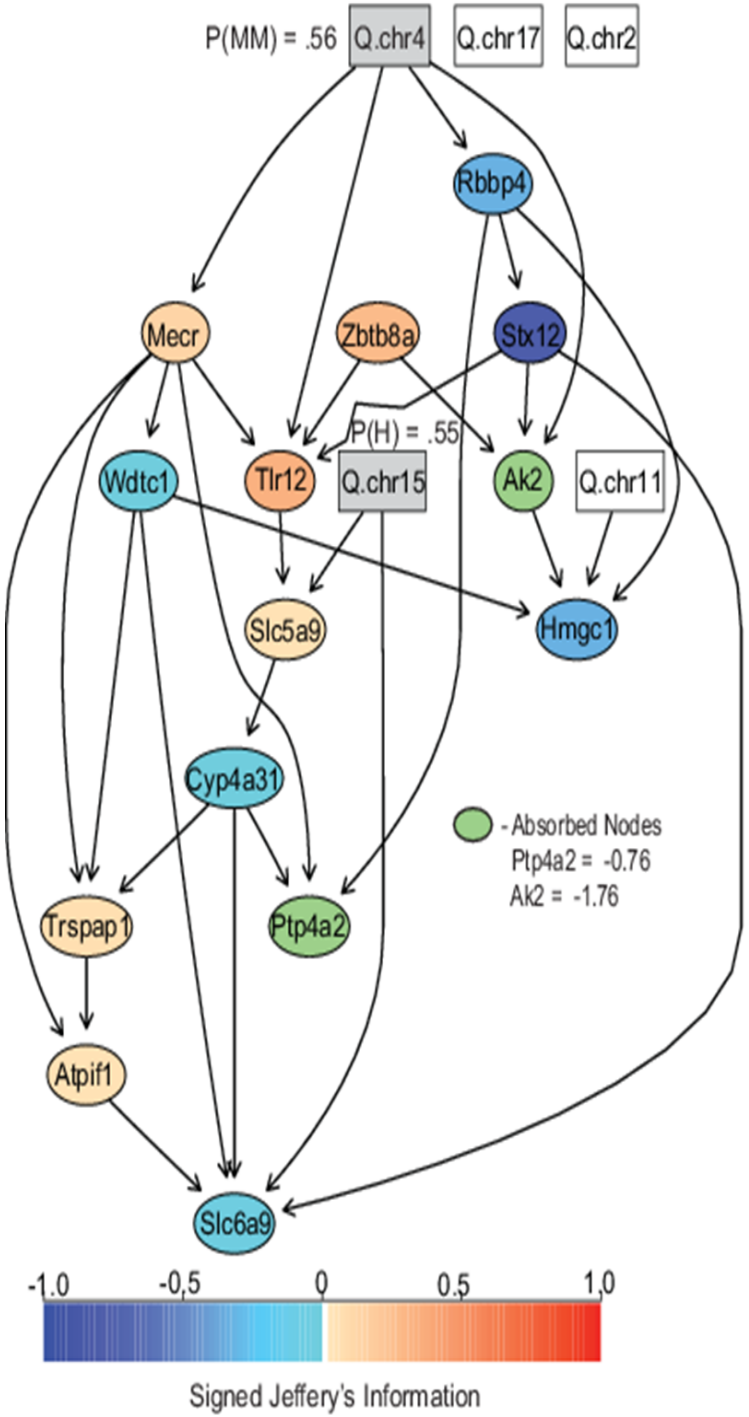
- **Decrease** in the marginal mean of ***Mecr*** indicating inhibition of fatty acid synthesis
- **Increase** in the marginal mean of ***Wdtc1*** which plays a role in negative regulation of fatty acid biosynthesis.
- **Inhibition** of sodium dependent glucose transport f ***Slc5a9***
- **Activation** of sodium and chloride dependent glycine transport ***Slc6a9***



Results: Two lines of evidence







Conclusions

- Belief propagation, which enables computational *in silico* predictions of the system-wide response inhibition or activation of phenotypes (perturbation(s)).

- Applications reveal coordination and co-regulation between sub-pathways in response to perturbation(s) of phenotypes in the network. This information is not revealed through network topology alone.

- A first step toward alleviating longstanding issues associated with model interpretation of genotype-phenotype networks.

- Insights provide a new layer of information, which may drive hypotheses generation, and the development of new experiments.

- Promising avenue for integration of probabilisitic constraints into a deterministic steady-state cellular model.

References

1. Lauritzen SL (1992) Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association 87: 1098-1108.

2. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences 186: 453-461.

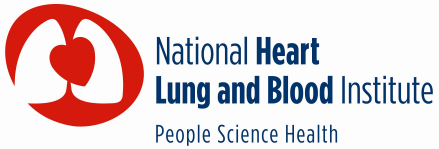
3. Rockman MV (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. Nature : 738 - 744.

Acknowledgements

Janhavi Moharil

Gary Churchill

DMS-1312250



F32 Ruth L. Kirschstein Fellowship
HL095240-01