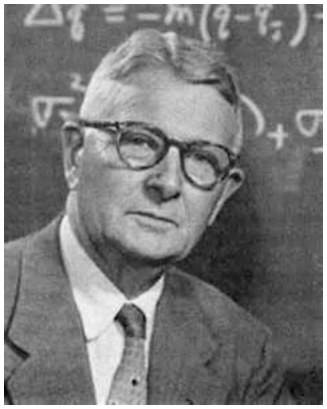
Lecture 10
Graphical Models in Quantitative Genetics and Genomics

Guilherme J. M. Rosa

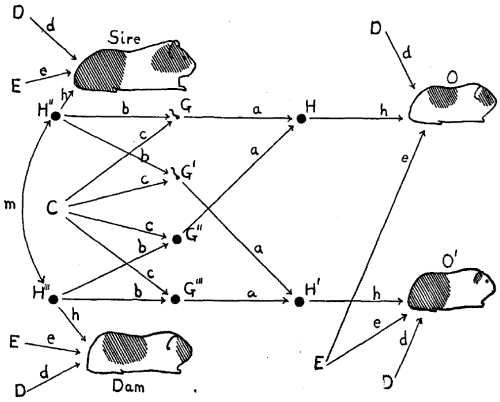
University of Wisconsin-Madison

Introduction to Quantitative Genetics
SISG, Seattle
15 – 17 July 2019

Path Analysis

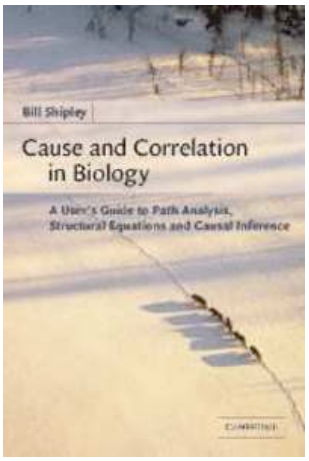


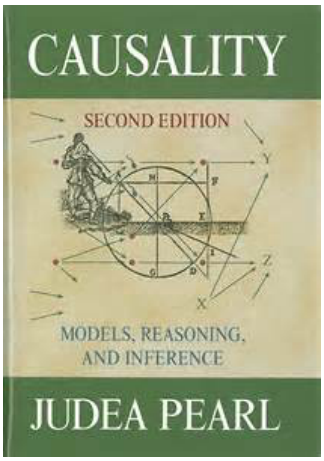
Sewall Wright
(1889-1988)

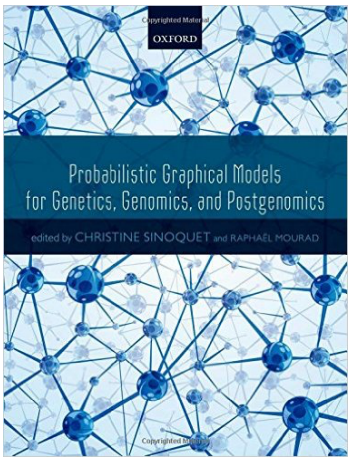


$h^2 + d^2 + e^2 = 1$

(Wright, 1921)

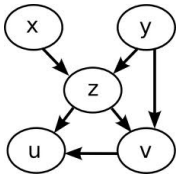






Outline

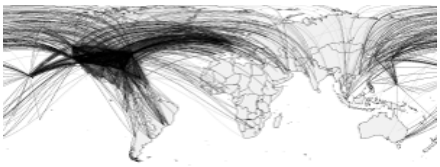
- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal inference
  - Visualization and model selection tool
- Concluding Remarks



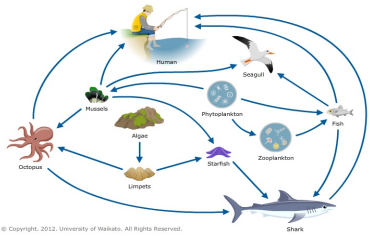
Networks

Social networks

Computer networks

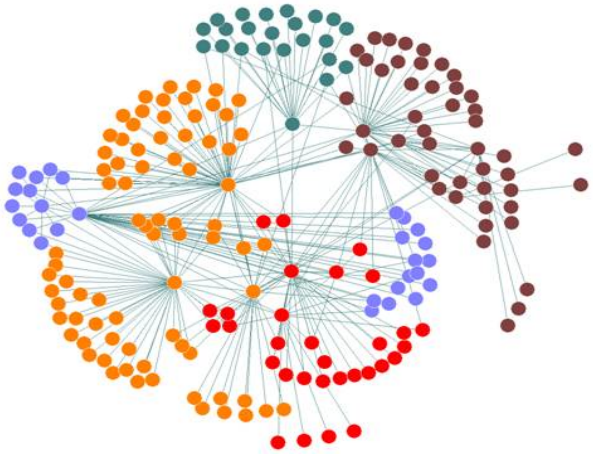


Airport hubs



Food web

Gene Networks

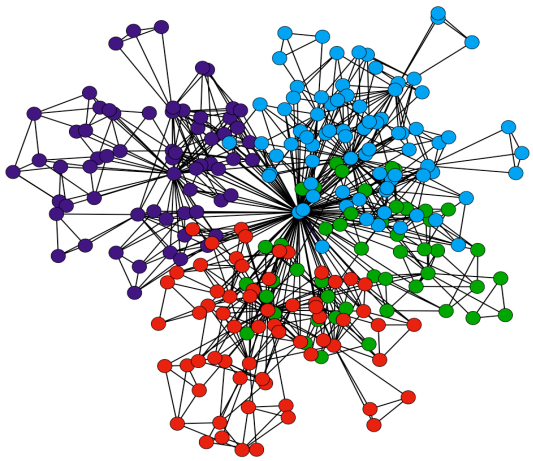


Gene regulation networks, co-expression, epistasis networks, etc.

Correlation Networks

- Attempt to systems understanding of organisms (understanding biology, e.g. master regulators, pharmaceutical intervention, etc.)
- Define some network parameter and topology features:
  - Degree (k_i) of node i: number of connections (degree centrality)
  - Communities and cliques; motifs
  - Betweenness of node i: measure of the number of shortest paths passing through node i (betweenness centrality)
  - Clustering coefficient: interconnectivity of nodes interacting with node of interest, $C_i = 2n_i/[k_i(k_i - 1)]$
- Compare topologies across genotypes, developmental stages, or environmental stress conditions
- Marginal or Partial Correlations

Marginal and Partial Correlations, and *d*-Separation



Marginal and Partial Correlations

- Pearson correlation coefficient: $\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$

Inference:

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2 \sum_{i=1}^{n}(y_i - \hat{\mu}_Y)^2}} \rightarrow t = \hat{\rho}_{XY}\sqrt{\frac{n-2}{1-\hat{\rho}_{XY}^2}} \sim t_{(\varphi=n-2)}$$

Fisher transformation:

$$z = \operatorname{arctanh}(\hat{\rho}_{XY}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_{XY}}{1-\hat{\rho}_{XY}}\right) \rightarrow \sqrt{n-3} \times |z| \overset{H_0:\rho=0}{\sim} N(0,1)$$

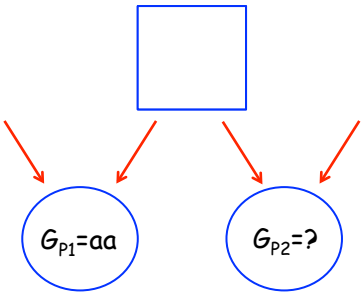
- Partial correlation: $\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}$

Conditioning on multiple variables: $P = \Sigma^{-1} \rightarrow \rho_{y_i y_j \cdot others} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}$

Marginal and Conditional Independence

Z → X → Y
Z ← X ← Y
Z ← X → Y

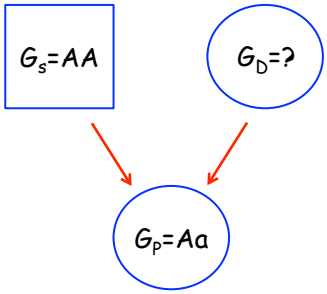
- ✓ Z & Y marginally not independent
- ✓ Conditioned on X they become independent



$Pr(G_{P2}) \neq Pr(G_{P2}|G_{P1})$

$Z \rightarrow X \leftarrow Y$

- ✓ Z & Y marginally independent
- ✓ Conditioned on X they are not independent
- ✓ Concept of collider, V-structure



$Pr(G_D=aa|G_S=AA) = Pr(G_D=aa) = q^2$

$Pr(G_D=aa|G_S=AA,G_P=Aa) = q \quad [\neq Pr(G_D=aa|G_P=Aa)]$

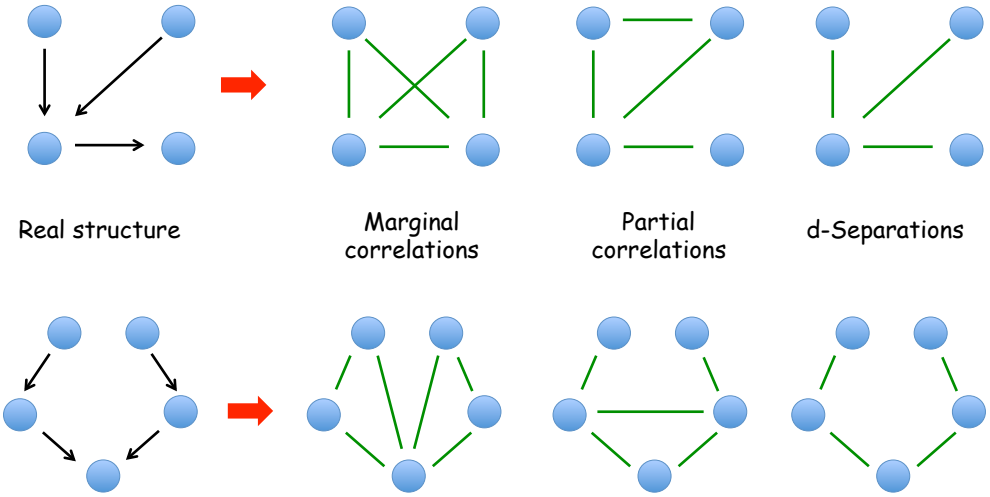
'Directed' Separation

⇨ *d-Separation* concept:

Two variables X and Y are said to be d-separated by Q if there is no active path between any X and Y conditionally on Q

(Verma and Pearl 1988, Pearl 1998, Geiger et al. 1990)

Examples

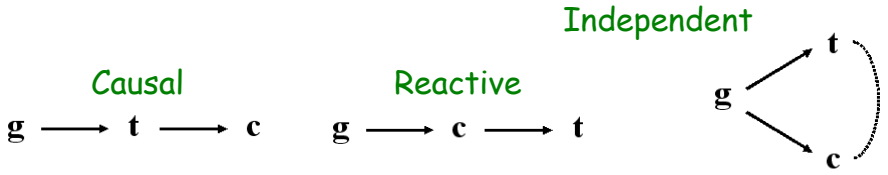


Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal inference
  - Visualization and model selection tool
- Concluding Remarks

Applications: Flow of Information from DNA to Phenotype

- Example with 3 nodes (Schadt et al. 2005): polymorphism (g), expression (t) and disease outcome (c)
- Causal, reactive and independent models:

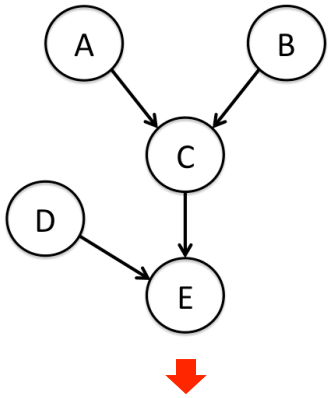


- Likelihood-based causality model selection (LCMS):

$$\begin{cases} C: p(g,t,c) = p(g)p(t \mid g)p(c \mid t) \\ R: p(g,t,c) = p(g)p(c \mid g)p(t \mid c) \\ I: p(g,t,c) = p(g)p(t \mid g)p(c \mid g,t) \end{cases}$$

Bayesian Networks

- Graphic representation of a probability distribution over a set of variables → DAG



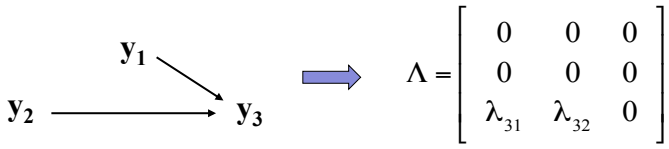
- Parents: Nodes that are directed to another node(s)
- Child: The descendent node
- Spouse: A node is defined as a spouse when it shares a child with another node

$$\Pr(A,B,C,D,E) = \Pr(E \mid C,D)\Pr(C \mid A,B)\Pr(D)\Pr(B)\Pr(A)$$

Inference Steps

① Structure Learning

- Score-based algorithms
- Constraint-based algorithms



① Parameter Estimation

$$\mathbf{y} = \Lambda\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Maximum Likelihood or Bayesian Inference

Structure Learning

- Constraint-based algorithms
  - IC, PC – Spirtes et al. (2001)
  - Grow-Shrink (GS) – Margaritis (2003)
  - Incremental Association Markov Blanket (IAMB) – Tsamardinos et al. (2003)
  - Max-Min Parents & Children (MMPC)
- Score-based algorithms
  - Hill Climbing (HC) – Bouckaert (1995)
  - Tabu Search (Tabu)
- Hybrid structure learning algorithms
  - Sparse Candidate (SC) – Friedman et al (1999)
  - Max-Min Hill Climbing (MMHC) – Tsamardinos et al. (2006)

Constraint-based algorithms

- Series of conditional independence tests (parametric, semiparametric and permutation)
  - Linear correlation or Fisher's Z (continuous data; multivariate normal distribution)
  - Pearson's X^2 or mutual information (categorical data; multinomial distribution)
  - Jonckheere-Terpstra (ordinal data)

Score-based algorithms

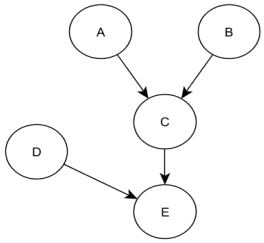
- Different score functions
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - multinomial log-likelihood, Dirichlet posterior density (BDe) or K2 score (categorical data)

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal inference
  - Visualization and model selection tool
- Concluding Remarks

Applications: Parsimonious Models for Multi-Trait Analysis

- k traits (nodes) → k(k - 1)/2 covariances
- Matrix Λ of SEM potentially with fewer parameters
- Model comparison using traditional techniques such as AIC, BIC, DIC etc.
- Example:



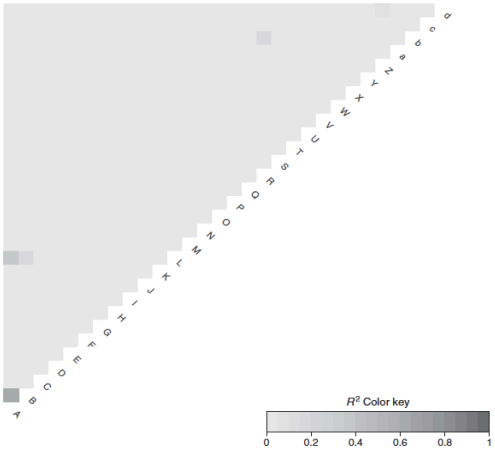
- Structure matrix Σ with 10 covariance parameters
- Matrix Λ with 4 unconstrained parameters

Example: Multilocus Linkage Disequilibrium

- 4898 Holstein bulls genotyped with the Illumina BovineSNP50 Bead Chip, provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD, USA)
- 36 778 SNP markers after editing and imputation using fastPHASE 1.4.0
- BN using SNPs with highest effects (BLasso)
- Tabu search algorithm with BDe scoring metric, and IAMB algorithm with X^2 test

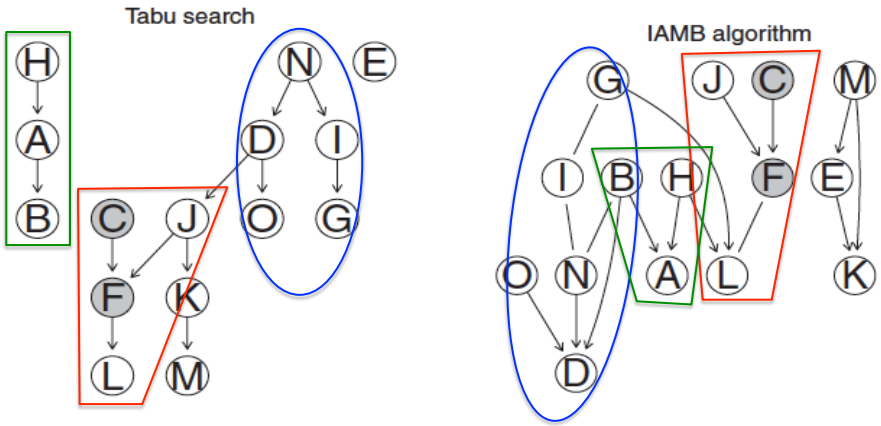
Morota G, Valente BD, Rosa GJM, Weigel KA and Gianola D. An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. *J. Anim. Breed. Genet.* 129: 474-487, 2012.

Pairwise LD among SNPs (r^2)

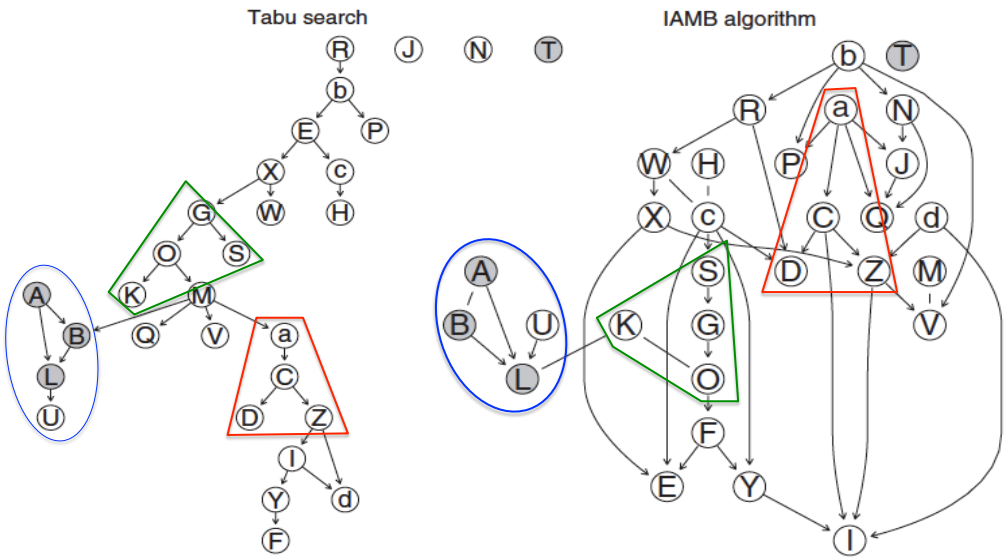


LD among the top 30 SNPs with the largest absolute posterior means using the r^2 metric.

Bayesian LD network



BNs learned by the Tabu and the IAMB searches for the 15 SNPs with the largest absolute posterior means.
Grey-filled nodes are SNPs located in chromosome 14.

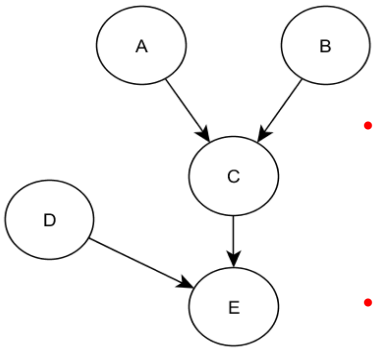


Similarly as before, for the 30 SNPs with the largest absolute posterior means.

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal inference
  - Visualization and model selection tool
- Concluding Remarks

Applications: Prediction, Markov Blanket



- Markov Blanket (MB): a MB of a node is defined as the set containing its parent(s), child(ren) and spouse(s)
- Conditionally on its MB, a node is independent from all other nodes

Examples: MB(D) = {C,E}; MB(E) = {C,D}; MB(C) = {A,B,D,E}

Example: Egg Production in Poultry

- Two strains (L1 and L2) of European Quail
- 31 traits (female quails):
  - Body weight
  - Weight gain
  - Age at first egg
  - Egg production
  - Egg quality traits

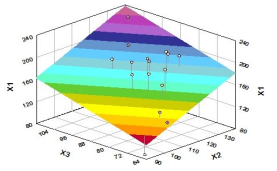
Felipe VPS, Silva MA, Valente BD and Rosa GJM. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry Science* 94(4): 772-780, 2015.

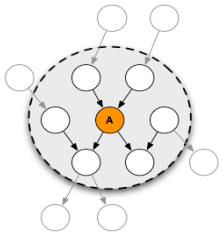
Material & Methods

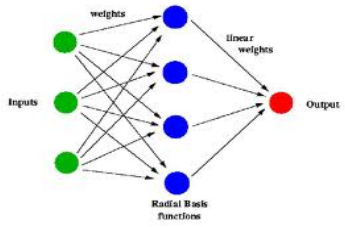
- Sample sizes (training and test sets):
  - Line 1 (90 + 90), Line 2 (102 + 103)
- Traits:
  - Weekly body weight (birth to 35 d, BW1 to BW6)
  - Weight gain (0-35 and 21-35 d, WG1 and WG2)
  - Age at first egg (AFE)
  - Egg quality traits, four time points: 125, 170, 215, 260 d
    Egg Weight – Ew, Yolk Weight – Y, Egg Shell Weight – ES
    Egg White Weight – EW, Egg Specific Gravity - DENS
  - Partial Egg Production (35-80d, EP1) and
    Total egg production (35-260d, TEP)

Material & Methods

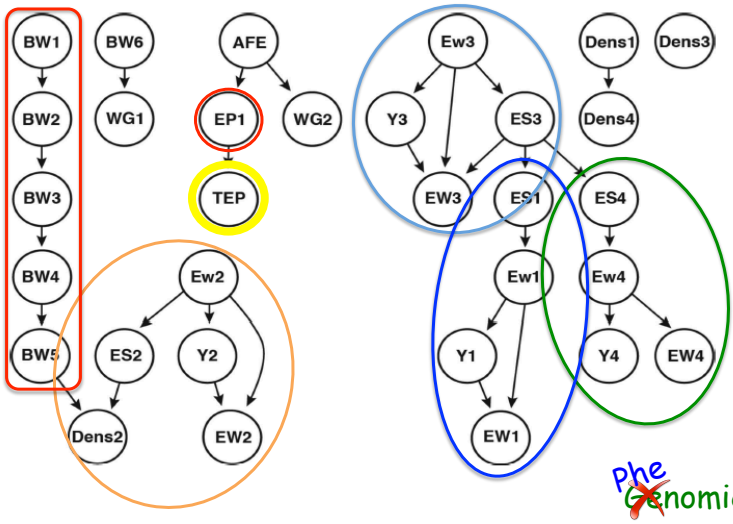
- Multiple regression analysis
  - Step-wise OLS
- Bayesian Networks
  - MB detection
- Artificial Neural Networks
  - Machine learning tool to map relationship between inputs and output



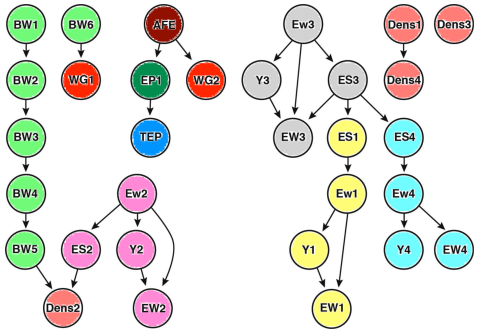




- Structure Learning (L1): Given EP1, TEP is independent from the other traits

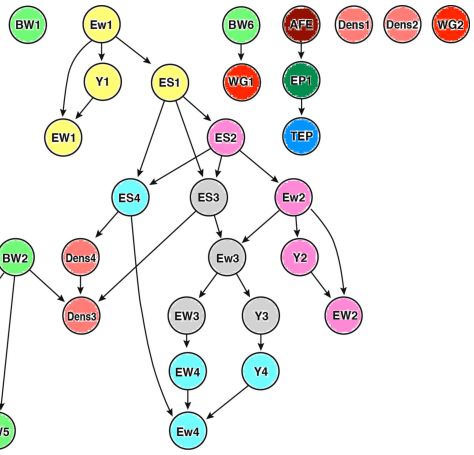


phe
Genomic Prediction!

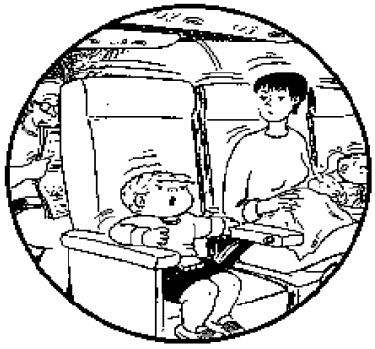


L1 Structure

L2 Structure



Causal Inference



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

Association vs. Causation

z ------- y

"z is associated with y"

z → y

"z causes y"

alcohol and drunkenness

z ← y

"y causes z"

police officers and criminality

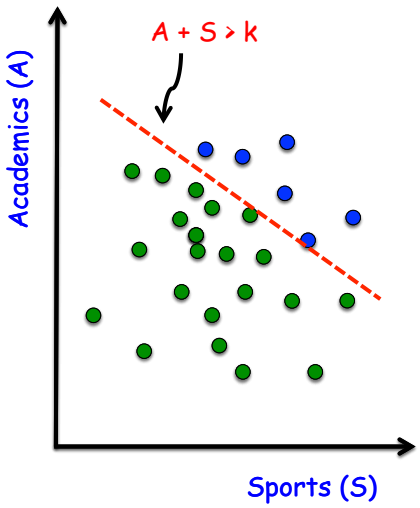
x → z, x → y

"x causes z & y"

water consumption and dehydration

Confounders

Selection Bias



Alcohol & Fats

It's a relief to know the truth after all those conflicting medical studies.

The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.

The French eat a lot of fat and also suffer fewer heart attacks than the British or Americans.

The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans.

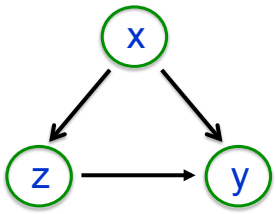
The Italians drink excessive amounts of red wine and also suffer fewer heart attacks than the British or Americans.

The Germans drink a lot of beer and eat lots of sausages and fats and suffer fewer heart attacks than the British or Americans.

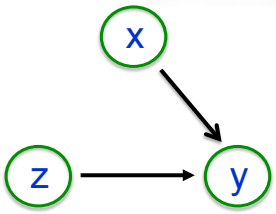
Conclusion: Eat and drink what you like. Speaking English is apparently what kills you.

Randomized Experiments

⇨ Testing the effect of z on y



Causal relationship between variables



Effect of randomization applied to variable z

Observational Studies

- ⇨ Lack of randomization due to legal, ethical, or logistics reasons
- ⇨ Potential bias and confounding effects
- ⇨ Example:
  Parenthood and life expectancy

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal inference
  - Visualization and model selection tool
- Concluding Remarks

The IC Algorithm
(Inductive Causation; Verma and Pearl 1991)

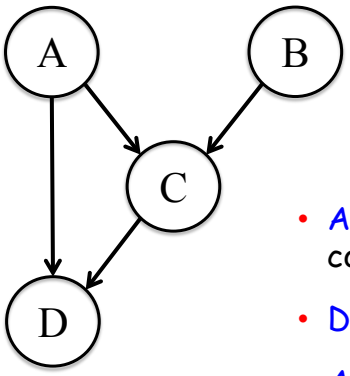
Step 1: Undirected graph (search for d-separations; connect adjacent variables)

Step 2: Partially oriented graph (search for colliders)

Step 3: Attempt to orient remaining undirected edges such that no new colliders or cycles are generated

Step 1: skeleton; Step 2: V structures

Applications: Causal Inference



- Arrows: Causal interpretation; consequences of intervention
- Direct, indirect and total effects
- Additional assumptions: Markov condition, faithfullness and causal sufficiency assumptions

Applications: Causal Inference

- Prediction of the result of an intervention (gene knockout, management decision, treatment effect)
- Estimation of causal effects:

  If the causal DAG is known and the distribution is multivariate Gaussian, then the causal effect (β) of X on Y can be estimated from the regression :

  $$E[Y] = m + \beta X + pa(X)$$

- i.e., DAG determines adjustment variables
  [backdoor adjustment; Pearl (1993)]

GENETICS | GENOMIC SELECTION

The Causal Meaning of Genomic Predictors and How It Affects Construction and Comparison of Genome-Enabled Selection Models

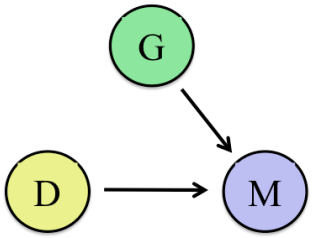
Bruno D. Valente,*,†,1 Gota Morota,† Francisco Peñagaricano,† Daniel Gianola,*,†,‡ Kent Weigel,* and Guilherme J. M. Rosa†,‡

*Departments of Dairy Science, †Animal Sciences, and ‡Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706

Genetics, Vol. 200, 483–494 June 2015

Example 1: Simulation Settings

- Consider the following causal network involving a
  - Genetics component (G)
  - Disease incidence (D), and
  - Milk yield (M)



- The following model was used to simulate data:

$$\begin{cases} y_D = \mu_D + e_D \\ y_M = \mu_M - 1.5 y_D + u_M + e_M \end{cases}$$

with

$$\mathrm{Var}\left[u_M\right] = 1.0$$

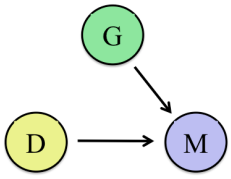
$$\mathrm{Var}\begin{bmatrix} e_D \\ e_M \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 1: Model Comparison

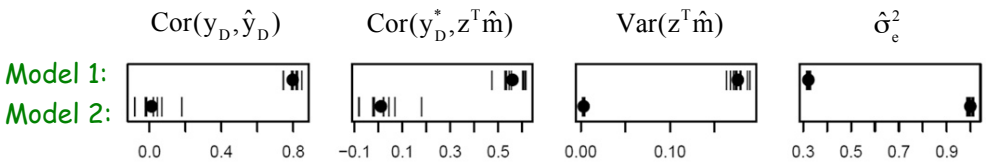
- Which model is better for the analysis of Disease?

Model 1: $y_D = \mu + \beta y_M + z^T m + e$

Model 2: $y_D = \mu + z^T m + e$

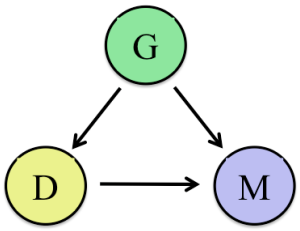


- Results:



Example 2: Simulation Settings

- Consider the following causal network involving a
  - Genetics component (G)
  - Disease incidence (D), and
  - Milk yield (M)



- The following model was used to simulate data:

$$\begin{cases} y_D = \mu_D + u_D + e_D \\ y_M = \mu_M - 1.5 y_D + u_M + e_M \end{cases}$$

with

$$\text{Var}\begin{bmatrix} u_D \\ u_M \end{bmatrix} = \begin{bmatrix} 0.3 & 0.25 \\ 0.25 & 1 \end{bmatrix}$$

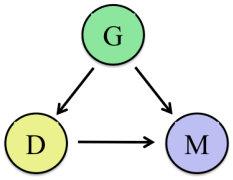
$$\text{Var}\begin{bmatrix} e_D \\ e_M \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 2: Model Comparison

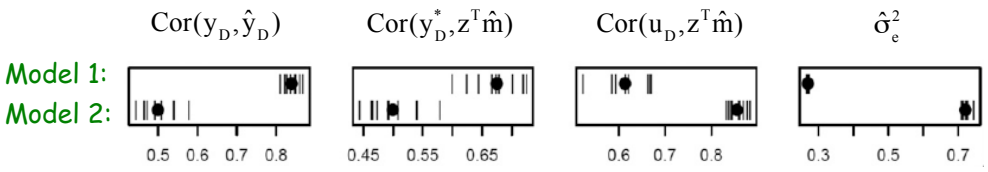
- Which model is better for the analysis of Disease?

Model 1: $y_D = \mu + \beta y_M + z^T m + e$

Model 2: $y_D = \mu + z^T m + e$

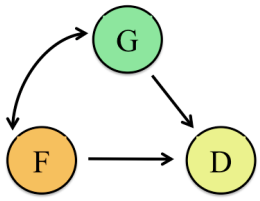


- Results:



Example 3: Simulation Settings

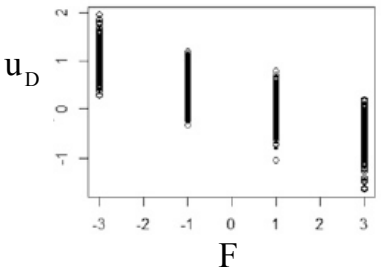
- Consider the following causal network involving a
  - Genetics component (G)
  - Farm effect (F), and
  - Disease incidence (D)



- The following model was used to simulate data:

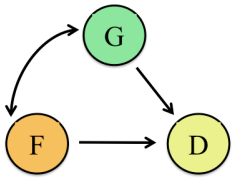
$$y_D = \mu_D + F + u_D + e_D$$

with $\mathrm{Var}[u_D] = 0.30$, $\mathrm{Var}[e_D] = 0.70$

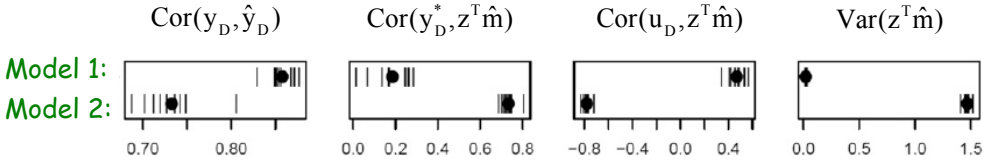


Example 3: Model Comparison

- Which model is better for the analysis of Disease?
  - Model 1: $y_D = \mu + F + z^T m + e$
  - Model 2: $y_D = \mu + z^T m + e$



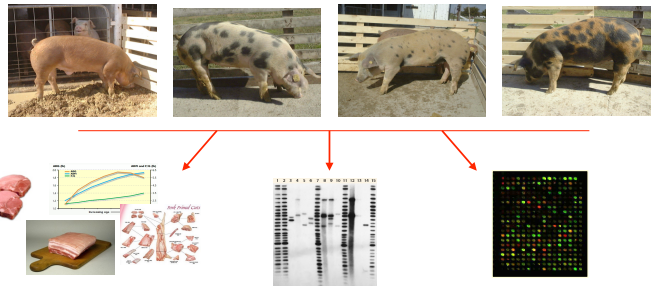
- Results:



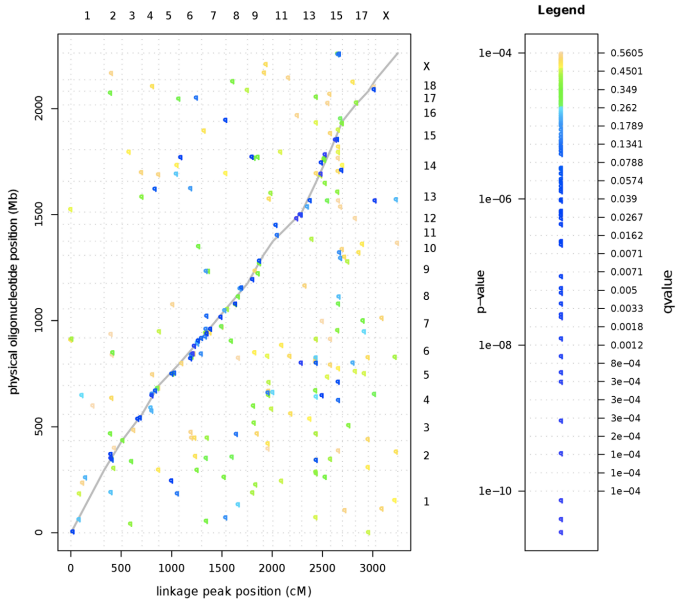
Example: Genetical Genomics

- F_2 Population (Duroc x Pietran)
- Genetical Genomics (eQTL Mapping)

1,000 F_2 progeny
25 growth traits
30 carcass traits
144 microsatellite loci
SNPs and Microarrays

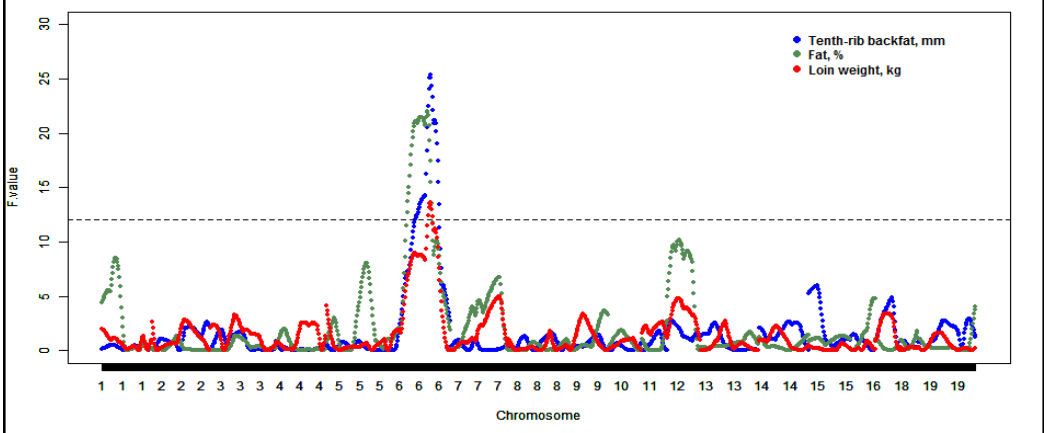


Steibel JP, Bates RO, Rosa GJM, et al. Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. *PLoS One* 6(2): e16766, 2011.

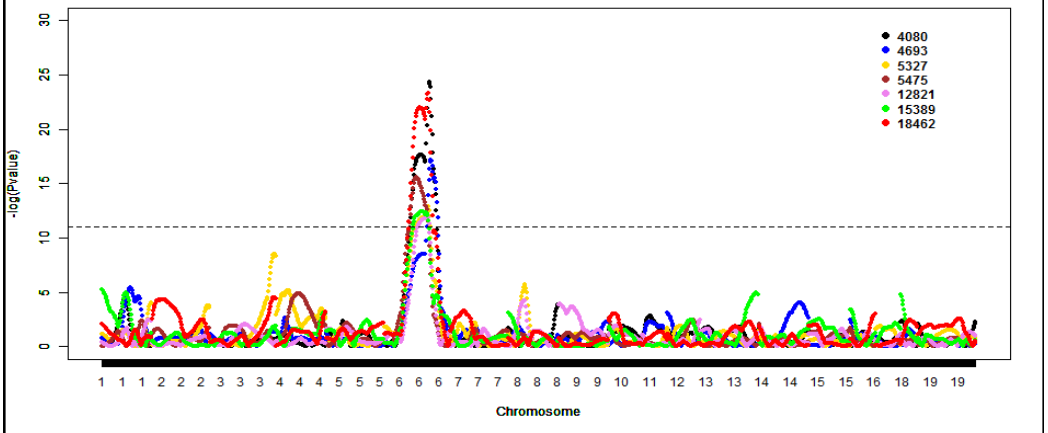


Global plot of physical position of oligonucleotide probe versus linkage position of eQTL across the pig genome. Points along the gray curve represent local eQTL (most likely cis-acting), while points off the line represent trans-acting eQTL.

pQTL Mapping Chr6



eQTL Mapping Chr6

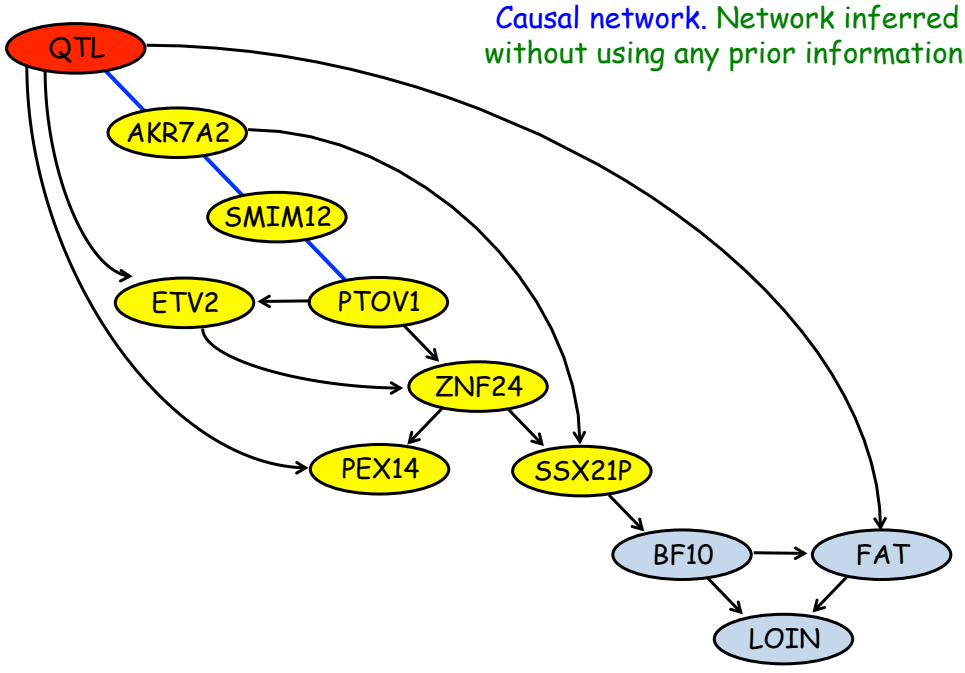


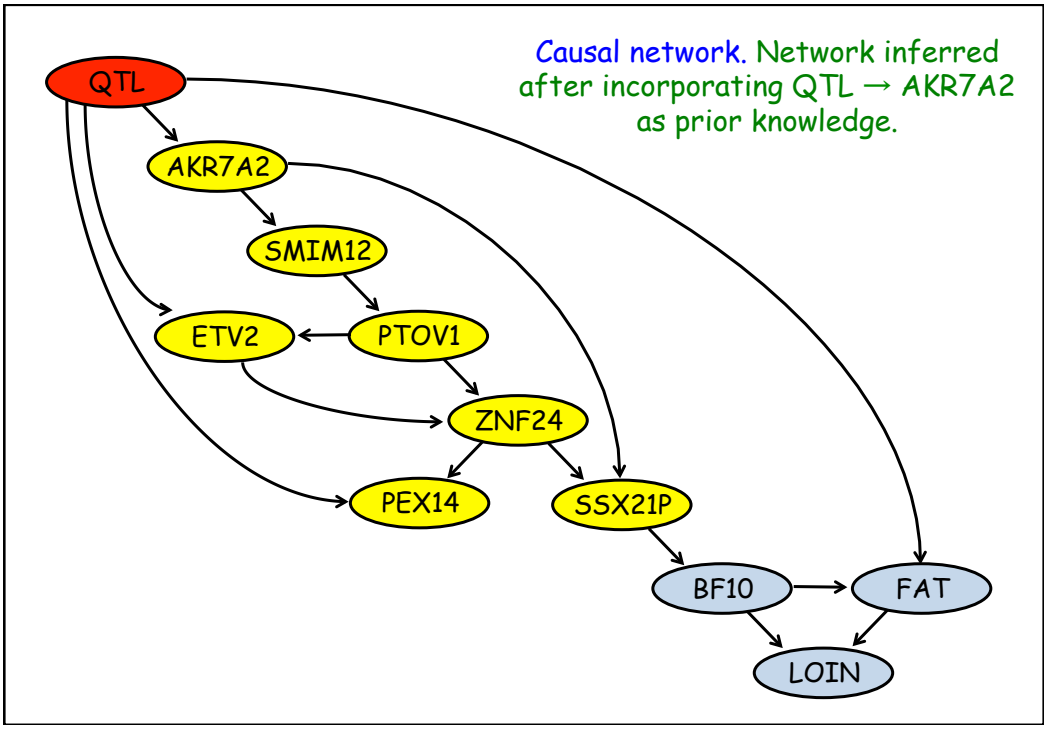
Exploring Causal Networks

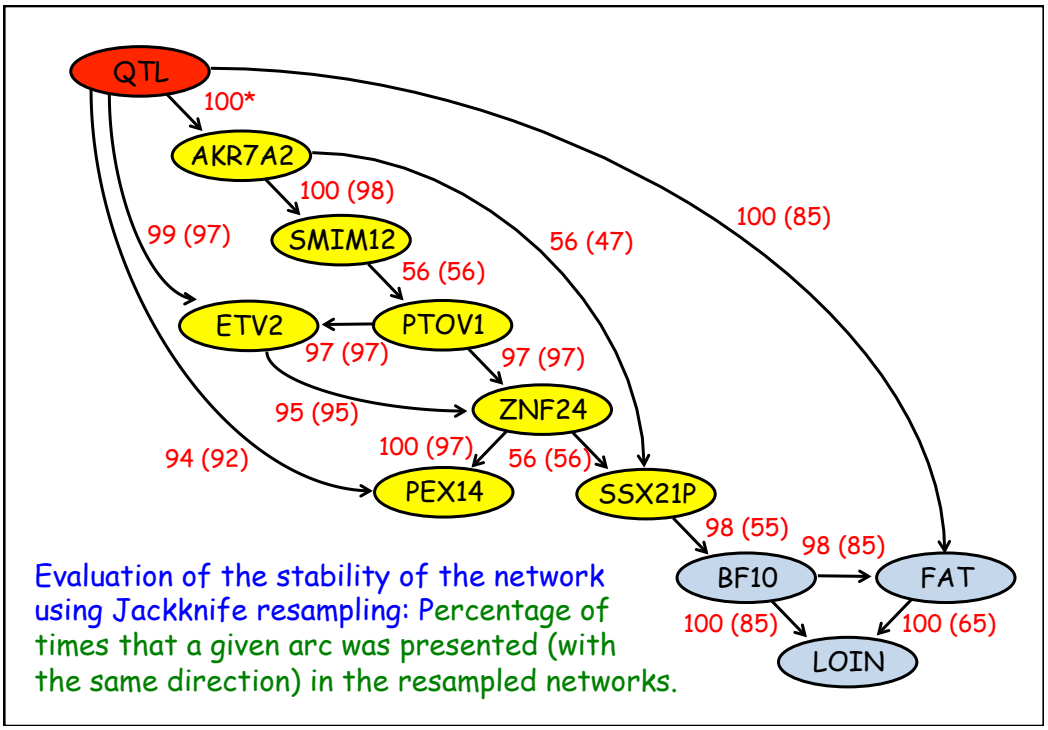
- Incremental Association Markov Blanket (IAMB) algorithm (Tsamardinos et al. 2003)
- *bnlearn* R package (Scutari 2010)
- Jackknife resampling

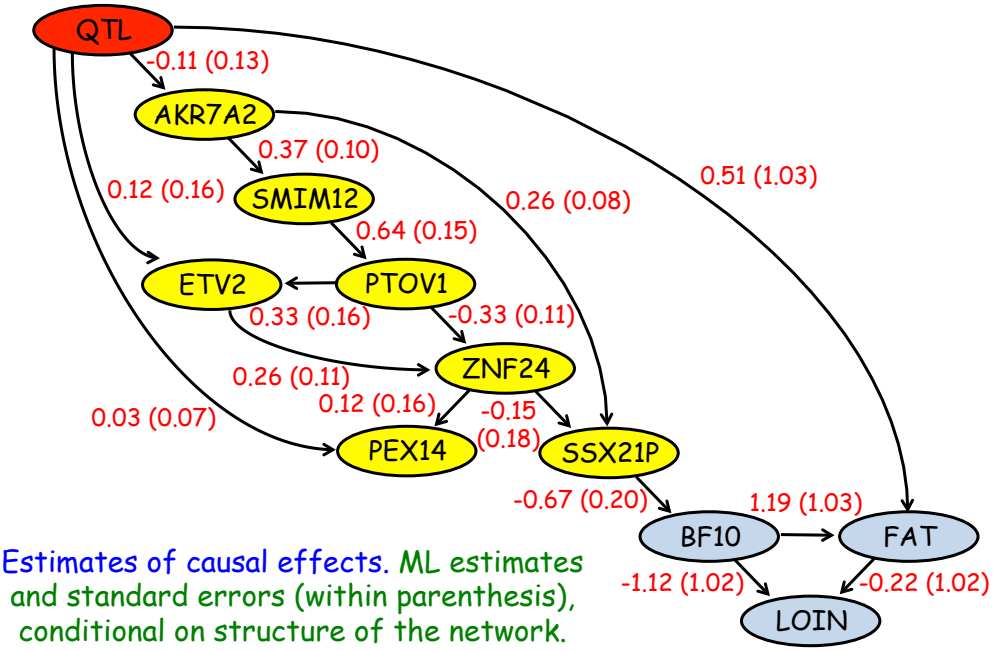
Peñagaricano F, Valente BD, Steibel JP, Bates RO, Ernst CW, Khatib H and Rosa GJM. Exploring causal networks underlying fat deposition and muscularity in pigs through the integration of phenotypic, genotypic and transcriptomic data. *BMC Systems Biology* 9:58, 2015.

Causal network. Network inferred without using any prior information.




Causal network. Network inferred after incorporating QTL → AKR7A2 as prior knowledge.
QTL
AKR7A2
SMIM12
ETV2
PTOV1
ZNF24
PEX14
SSX21P
BF10
FAT
LOIN


QTL
100*
AKR7A2
100 (98)
99 (97)
SMIM12
56 (56)
56 (47)
100 (85)
ETV2
PTOV1
97 (97)
97 (97)
95 (95)
ZNF24
94 (92)
100 (97)
56 (56)
PEX14
SSX21P
98 (55)
98 (85)
BF10
FAT
100 (85)
100 (65)
LOIN
Evaluation of the stability of the network using Jackknife resampling: Percentage of times that a given arc was presented (with the same direction) in the resampled networks.



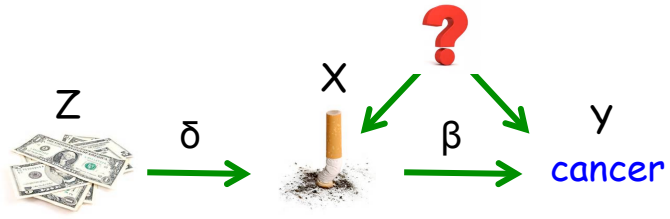
Estimates of causal effects. ML estimates and standard errors (within parenthesis), conditional on structure of the network.

Discussion

- ZNF24 encodes a member of the family of Krüppel-like zinc finger transcription factors and has critical roles in cell proliferation and differentiation
- The network model predicts that modulation of ZNF24 expression level should lead to a change in the expression of SSX2IP
- Recently, Li et al. (2009) evaluated potential ZNF24 target genes. For this purpose, the authors transiently overexpressed and silenced ZNF24 and then applied microarray assay in order to identify target genes.
- Notably, the overexpression of ZNF24 significantly decreased the expression of SSX2IP as predicted by our network. In addition, the silenced of ZNF24 resulted in a significant overexpression of SSX2IP (Li et al. 2009)

Li JZ, Chen X, Gong XL, Liu Y, Feng H, Qiu L, Hu ZL and Zhang JP. A transcript profiling approach reveals the zinc finger transcription factor ZNF191 is a pleiotropic factor. *BMC Genomics* 10, 2009.

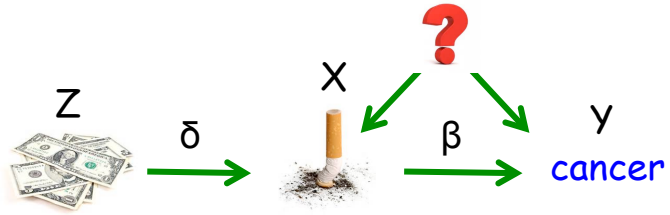
Instrumental Variable (IV)



$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta}_{IV} = (Z^T X)^{-1} Z^T Y$$

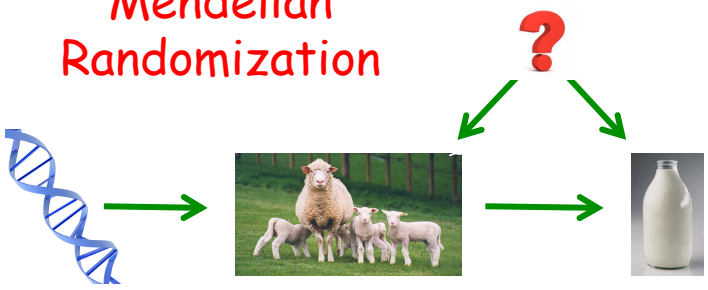
Instrumental Variable (IV)



Two-stage estimation:

1. Regress X on Z: $\hat{\delta} = (Z^T Z)^{-1} Z^T X$ and save predicted values $\hat{X} = Z\hat{\delta} = Z(Z^T Z)^{-1} Z^T X = P_Z X$
2. Regress Y on $\hat{X}$: $\hat{\beta}_{2SLS} = (X^T P_Z X)^{-1} X^T P_Z Y$

Mendelian Randomization



➔ Assumptions underlying the use of an instrumental variable (and the Mendelian randomization strategy):

- The instrumental variable (genotype) is associated with the 'exposure' of interest
- The genotype is independent of any confounding variable
- The association between genotype and outcome exists only because the genotype is associated with the exposure

Concluding Remarks

- Graphical Models: visualization/descriptive tool, prediction, causality, hypothesis generator
- Data driven + prior biological knowledge
- Causality inference: Markov condition, faithfullness and causal sufficiency assumptions
- Instrumental variable; Mendelian randomization

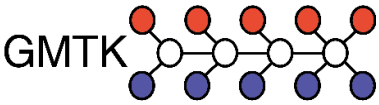
Rosa GJM, Valente BD, de los Campos G, Wu X-L, Gianola D and Silva MA. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution* 43: 6, 2011.

Software

bnlearn
deal
pcalg
catnet

The TETRAD Project

Causal Models and Statistical Data



The Graphical Models Toolkit