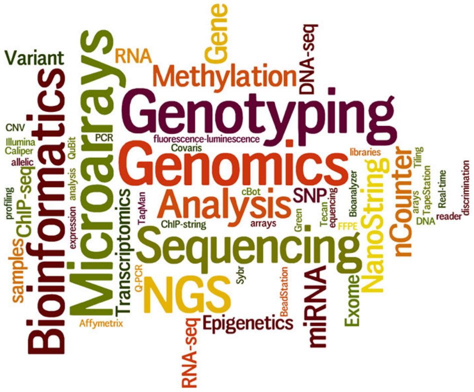
QTL Analysis and GWAS

Francisco Peñagaricano
University of Florida

High throughput technologies: omics data

- **genetic variants**
- **gene expression**
- **epigenetic modifications**
- **proteins** and **metabolites**



- measuring different **phenotypic traits**

unprecedented opportunities to uncover the **genetic architecture** underlying **phenotypic variation**

Main challenge:

decipher the flow of biological information

- integrate multiple sources of biological information in order to reveal the **causal biological networks** that underlie complex traits

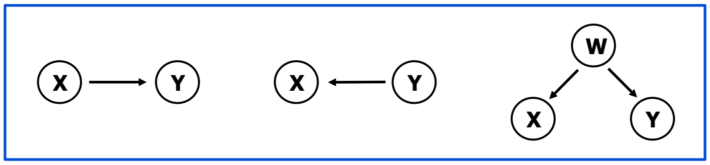
Why do we want to infer Causal Biological Networks?

- to better understand the biology of the traits
- to predict the behavior of complex systems
- to optimize management practices and breeding strategies

Causal Inference:

the idea is to infer **network structures** underlying a set of **correlated variables**

CORRELATION
≠
CAUSATION

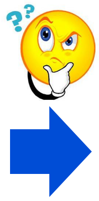


Causal Inference:

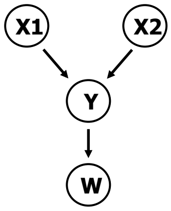
the idea is to infer **network structures** underlying a set of **correlated variables**

observed data

N	**X1**	**X2**	**Y**	**W**
1	x_{11}	x_{12}	y_1	w_1
2	x_{21}	x_{22}	y_2	w_2
..	..	..	..	..
n	x_{1n}	x_{2n}	y_n	w_n



causal structure

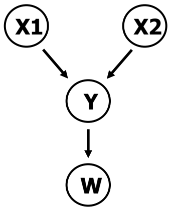


assumption: the pattern of conditional independencies observed in the data is compatible with the unknown causal model

Causal Inference:

the idea is to infer **network structures** underlying a set of **correlated variables**

true causal structure



observed pattern of conditional independencies

- **X1** and **X2** are marginally **independent**
- **X1** and **Y**/ **W** are marginally **dependent**
- **X2** and **Y**/ **W** are marginally **dependent**
- **Y** and **W** are marginally **dependent**

- Conditionally on **Y**, then **X1** and **X2** are **dependent**
- Conditionally on **Y**, then **X1/X2** and **W** are **independent**

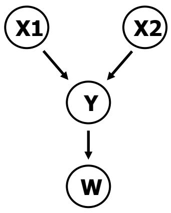
Causal Inference:

the idea is to infer **network structures** underlying a set of **correlated variables**

observed data

N	**X1**	**X2**	**Y**	**W**
1	x_{11}	x_{12}	y_1	w_1
2	x_{21}	x_{22}	y_2	w_2
..	..	..	..	..
n	x_{1n}	x_{2n}	y_n	w_n

causal structure

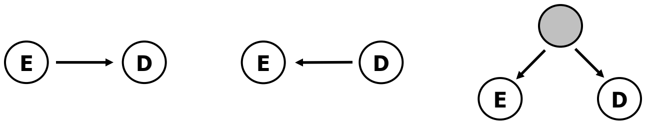


explore all causal hypotheses in order to find a causal model that is able to generate the observed pattern of cond. independencies

Genetics and Causal Inference

Motivation:

the expression (**E**) of gene is associated with a disease (**D**)



how can we infer the structure underlying this association?

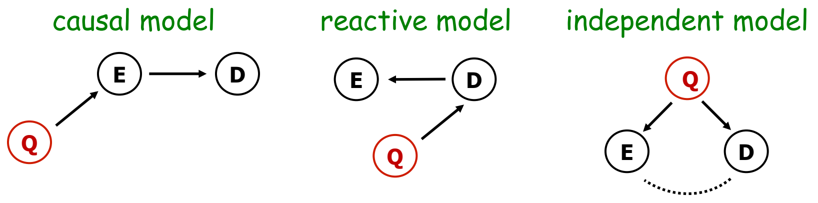
if **E** and **D** map to the same QTL, then we can use genetic information to infer the causal structure

Genetics and Causal Inference

Motivation:

the expression (**E**) of gene is associated with a disease (**D**)

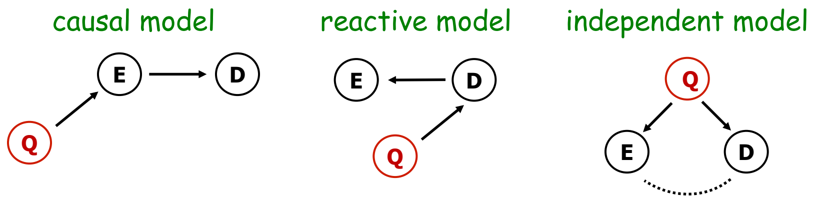
Assumption: **E** and **D** are controlled by a common **Q**



these models have distinct patterns of cond. independence (models are not distribution/likelihood equivalent)

Erick Schadt et al. (2005) Nat Genet. 37: 710-717

Genetics and Causal Inference



$$\text{model } \boldsymbol{C}\text{: } p(Q,E,D) = P(Q) \cdot P(E|Q) \cdot P(D|E)$$

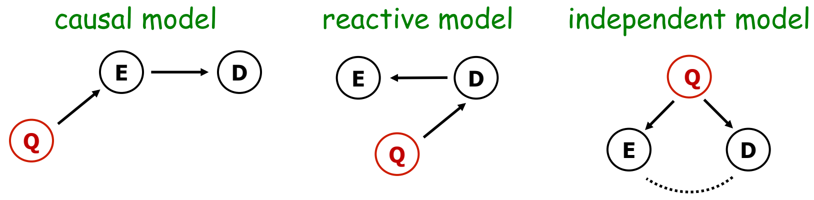
$$\text{model } \boldsymbol{R}\text{: } p(Q,E,D) = P(Q) \cdot P(D|Q) \cdot P(E|D)$$

$$\text{model } \boldsymbol{I}\text{: } p(Q,E,D) = P(Q) \cdot P(E|Q) \cdot P(D|Q,E)$$

these models have distinct patterns of cond. independence (models are not distribution/likelihood equivalent)

Erick Schadt et al. (2005) Nat Genet. 37: 710-717

Genetics and Causal Inference



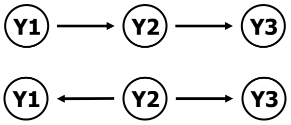
likelihood-based causality model selection

$$\hat{L} = P(x|\hat{\theta}, M)$$
$$AIC = -2 \cdot log(\hat{L}) + 2 \cdot k$$

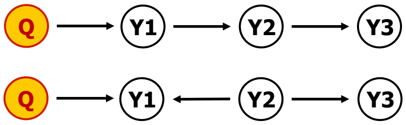
preferred model is the one with the **minimum AIC value**

Erick Schadt et al. (2005) Nat Genet. 37: 710-717

Multiple Phenotypes

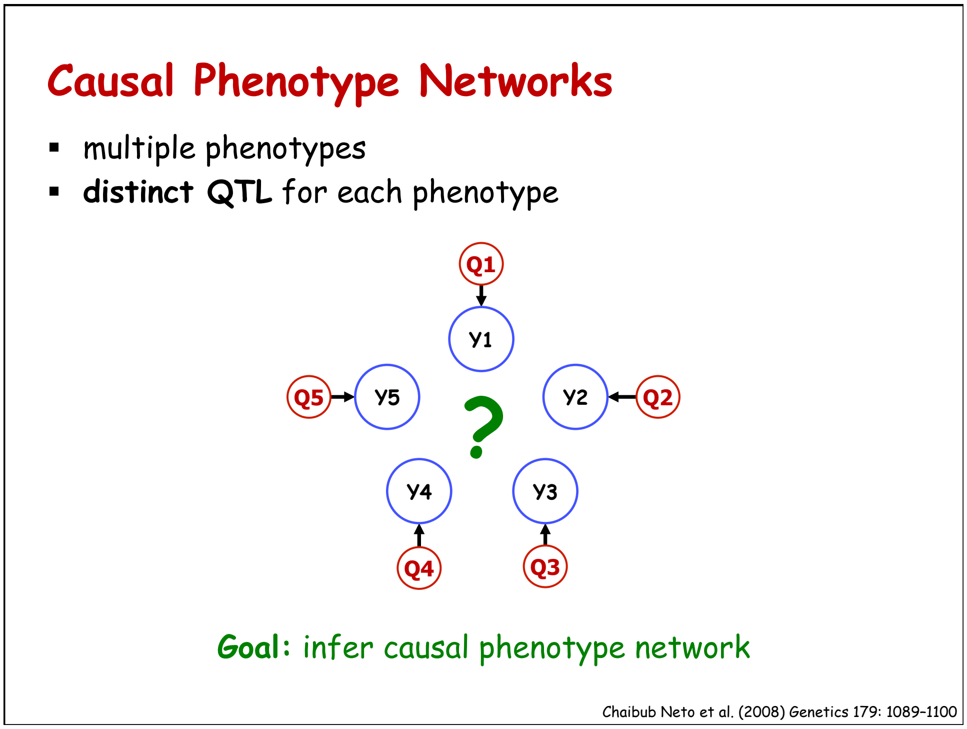


these models are **likelihood equivalent**



these extended models are no longer **likelihood equivalent**

adding **causal QTL nodes** to a phenotype network allows the inference of causal relationships between phenotypes

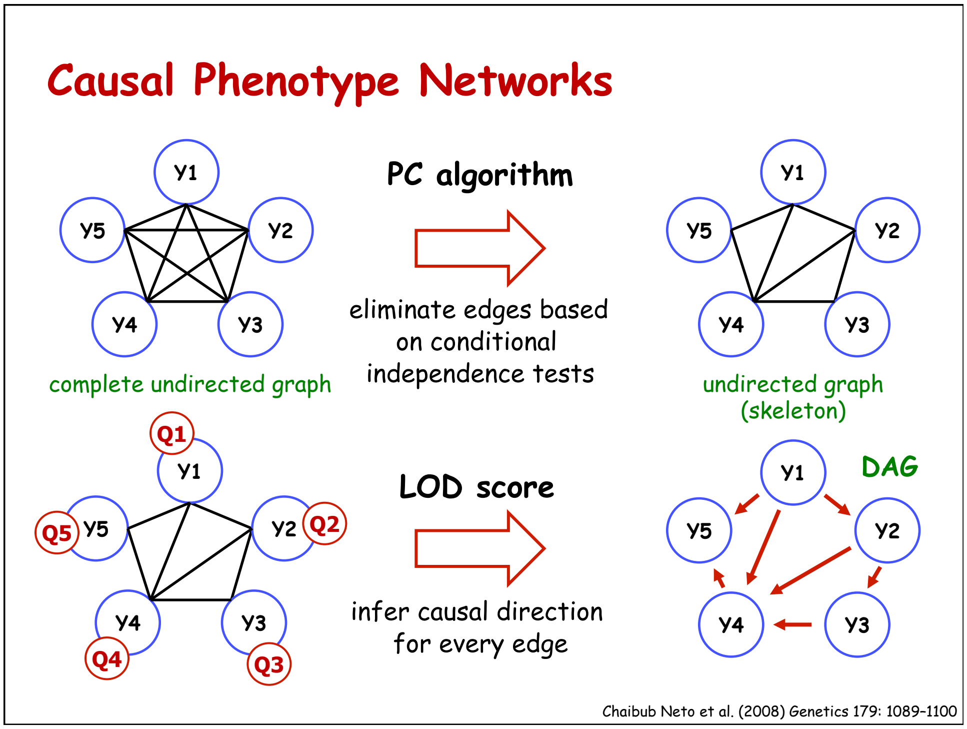


Causal Phenotype Networks

- multiple phenotypes
- **distinct QTL** for each phenotype

Goal: infer causal phenotype network

Chaibub Neto et al. (2008) Genetics 179: 1089-1100



Causal Phenotype Networks

complete undirected graph

PC algorithm

eliminate edges based on conditional independence tests

undirected graph (skeleton)

LOD score

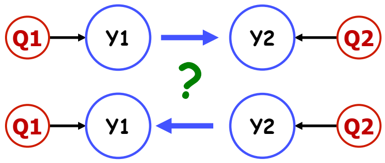
infer causal direction for every edge

DAG

Chaibub Neto et al. (2008) Genetics 179: 1089-1100

Causal Phenotype Networks

direction LOD score



$$LOD = log_{10}\left\{\frac{\prod_{i=1}^{n} f(y_{1i}|q_{1i})f(y_{2i}|y_{1i},q_{2i})}{\prod_{i=1}^{n} f(y_{2i}|q_{2i})f(y_{1i}|y_{2i},q_{1i})}\right\}$$

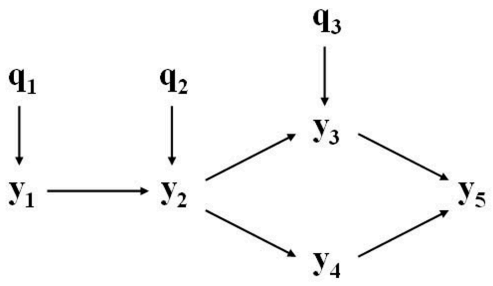
Causal Phenotype Networks

QDG algorithm

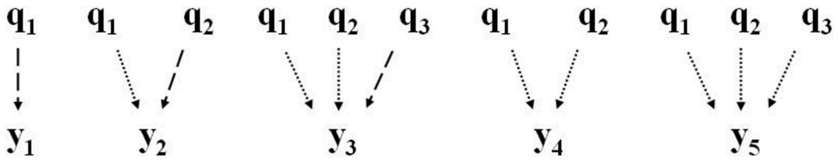
- QTLs are assumed to come from earlier gene mapping
- QTL mapping and net inference are performed separately

- poor estimation of QTL locations & effects may compromise the inference of phenotype networks
- ignoring causal phenotypes may bias mapping results by incorrectly **inferring QTLs** that have **indirect effects**

Causal Phenotype Networks



Single-trait QTL analysis



Rosa et al. (2011) Genet Sel Evol. 43: 6

Causal Phenotype Networks

QDG algorithm

- QTLs are assumed to come from earlier gene mapping
- QTL mapping and net inference are performed separately

- poor estimation of QTL locations & effects may compromise the inference of phenotype networks
- ignoring causal phenotypes may bias mapping results by incorrectly inferring QTLs that have indirect effects

Better Approach:

Joint inference of causal QTLs and causal network

Joint Inference QTLs & Network

Aim: perform joint inference of the genetic architecture and the causal phenotype network

- the genetic architecture should be inferred conditional on the phenotype network
- but the phenotype network is unknown ⋯

Solution: iterate between updating the genetic architecture and the phenotype network using a MCMC approach

Chaibub Neto et al. (2010) Ann Appl Stat. 4:320-339

Joint Inference QTLs & Network

QTLnet - Metropolis-Hastings algorithm

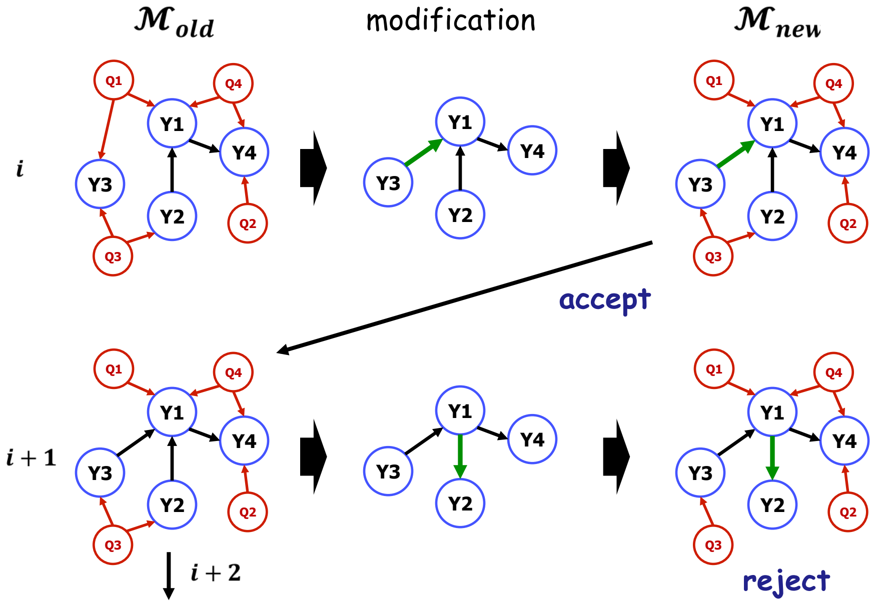
1. Propose a new phenotype network $\mathcal{M}_{new}$

   (by adding, deleting or reversing an edge from $\mathcal{M}_{old}$)
2. Recompute QTL locations and effects
3. Compute the marginal likelihood $\hat{p}(\mathbf{y}|\mathbf{q}, \mathcal{M}_{new})$
4. Accept $\mathcal{M}_{new}$ with probability

$$\alpha = min\left\{1, \frac{\hat{p}(\mathbf{y}|\mathbf{q}, \mathcal{M}_{new})p(\mathcal{M}_{new})q(\mathcal{M}_{old}|\mathcal{M}_{new})}{\hat{p}(\mathbf{y}|\mathbf{q}, \mathcal{M}_{old})p(\mathcal{M}_{old})q(\mathcal{M}_{new}|\mathcal{M}_{new})}\right\}$$

Chaibub Neto et al. (2010) Ann Appl Stat. 4:320-339

QTLnet algorithm



$\mathcal{M}_{old}$

modification

$\mathcal{M}_{new}$

i

accept

$i+1$

$i+2$

reject

Joint Inference QTLs & Network

output QTLnet algorithm: **it is not a single network**

Bayesian Model Averaging

Chaibub Neto et al. (2010) Ann Appl Stat. 4:320-339

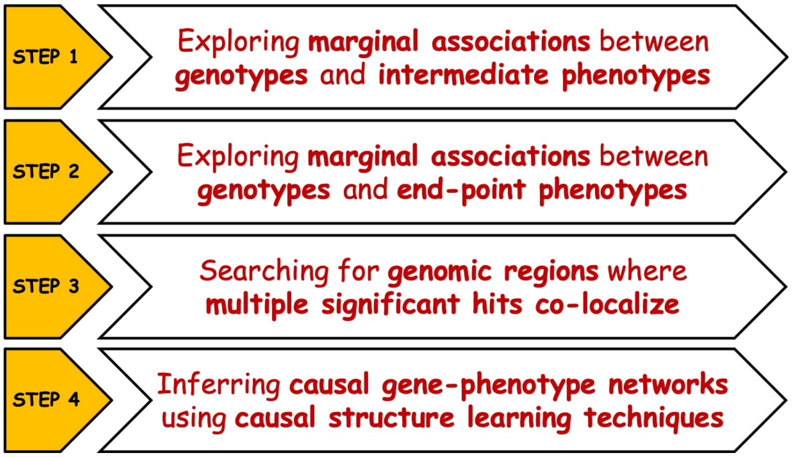
Integrating multi-omics data

multiple layers of information

- genetic variation
- gene expression
- epigenetic modifications
- proteins and metabolites
- phenotypic traits

Integrating multi-omics data

the goal is to reconstruct networks integrating **multiple layers of information**



STEP 1 — Exploring **marginal associations** between **genotypes** and **intermediate phenotypes**

STEP 2 — Exploring **marginal associations** between **genotypes** and **end-point phenotypes**

STEP 3 — Searching for **genomic regions** where **multiple significant hits co-localize**

STEP 4 — Inferring **causal gene-phenotype networks** using **causal structure learning techniques**

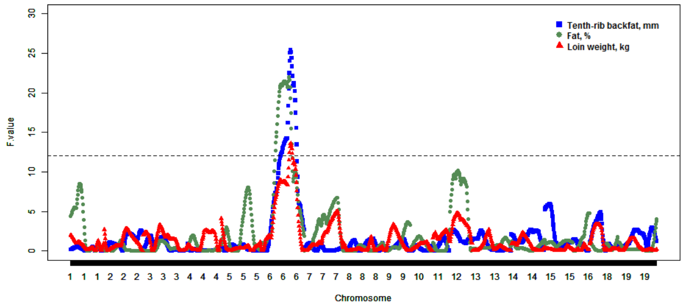
Peñagaricano et al. (2015) BMC Syst Biol. 9:58

Integrating multi-omics data

integrate **phenotypic**, **genotypic** and **transcriptomic data** from F_2 pig population

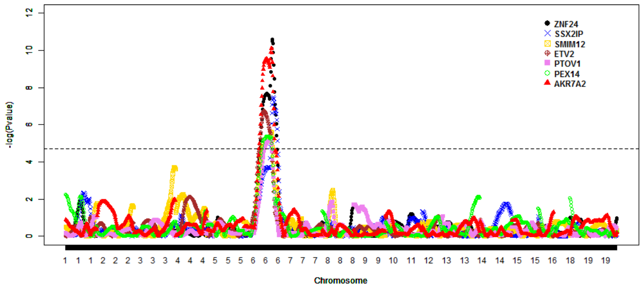
- several phenotypes for carcass traits
- genotypes for microsatellites spanning the whole genome
- gene expression data for almost 20,000 transcripts measured in loin muscle tissue

Integrating multi-omics data

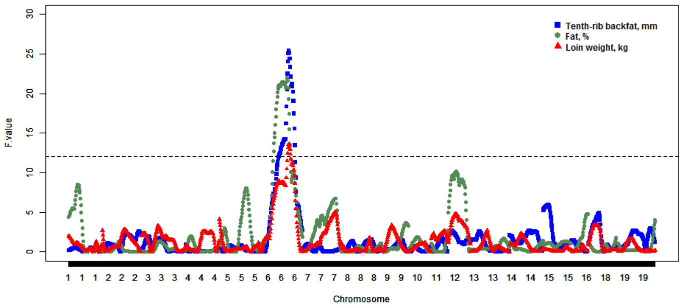


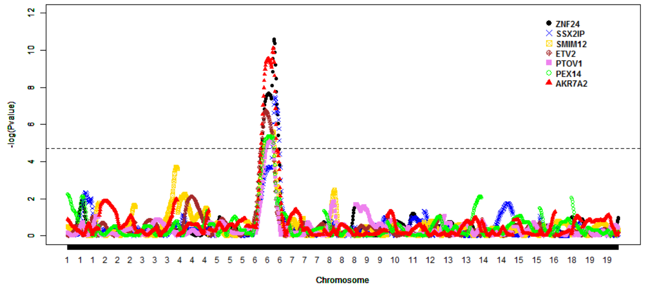
QTL mapping for phenotypic traits

Gene expression as a response variable (eQTL mapping)



can we infer causal links?



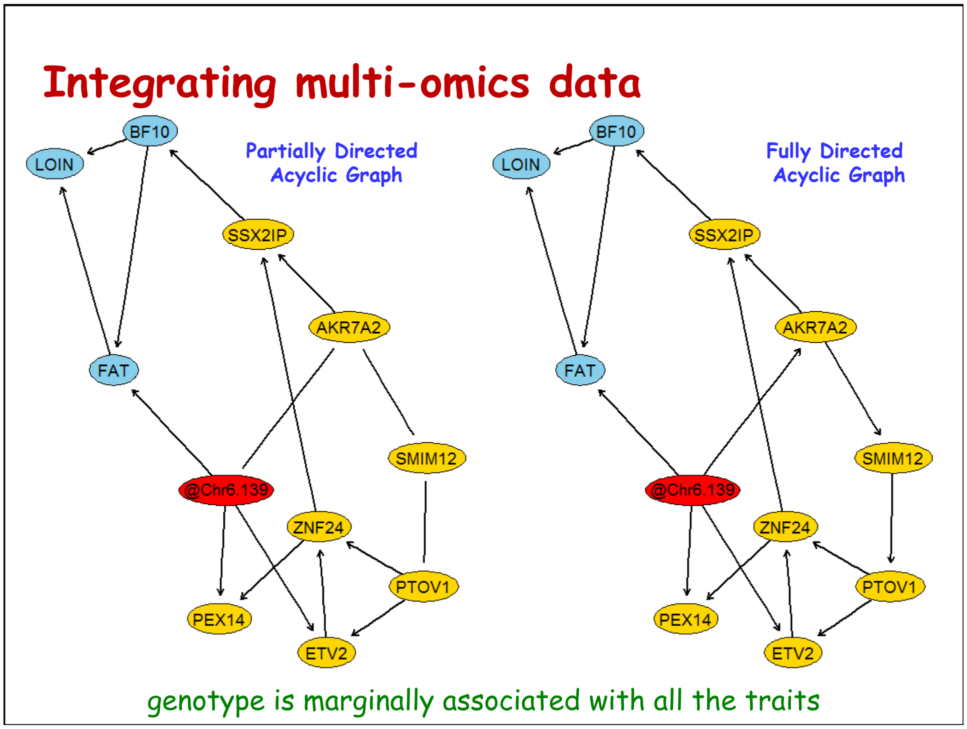


Causal Inference using IC Algorithm

1. for each pair of variables **X** and **Y**, search for set of other variables S_{XY} such that X and Y are independent given S_{XY}
   - if X and Y are **dependent** for every possible S_{XY}, then place an **undirected edge** between X and Y

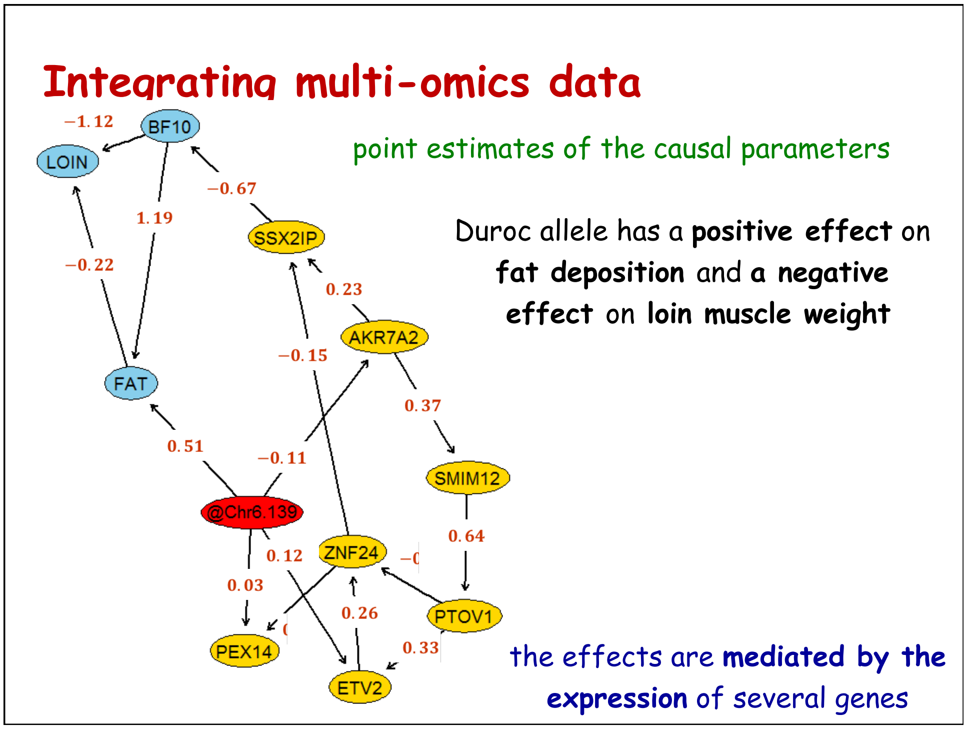
2. for each pair of non-adjacent variables X and Y with a common adjacent variable C, search for a set S_{XY} containing C such that X and Y are independent given S_{XY}
   - if there is no such set, then assign the direction of the edges X-C and C-Y as **X→C** and **C←Y**

3. in the partially directed graph, orient undirected edges without creating **new v-structures** or **directed cycles**



Integrating multi-omics data

genotype is marginally associated with all the traits

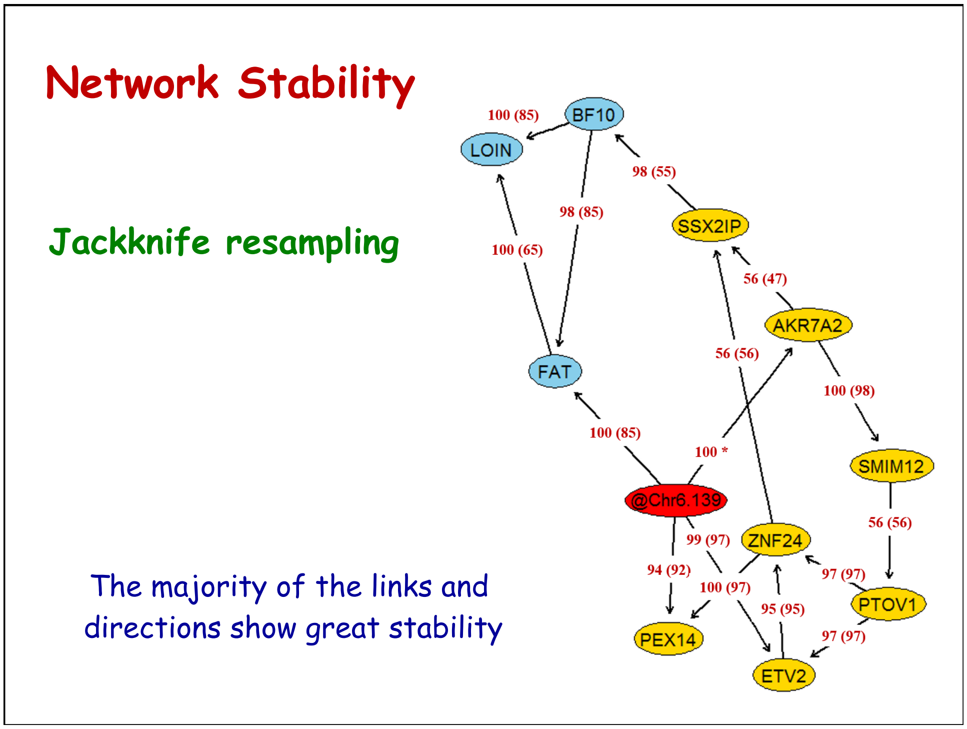


Integrating multi-omics data

point estimates of the causal parameters

Duroc allele has a **positive effect** on **fat deposition** and **a negative effect** on **loin muscle weight**

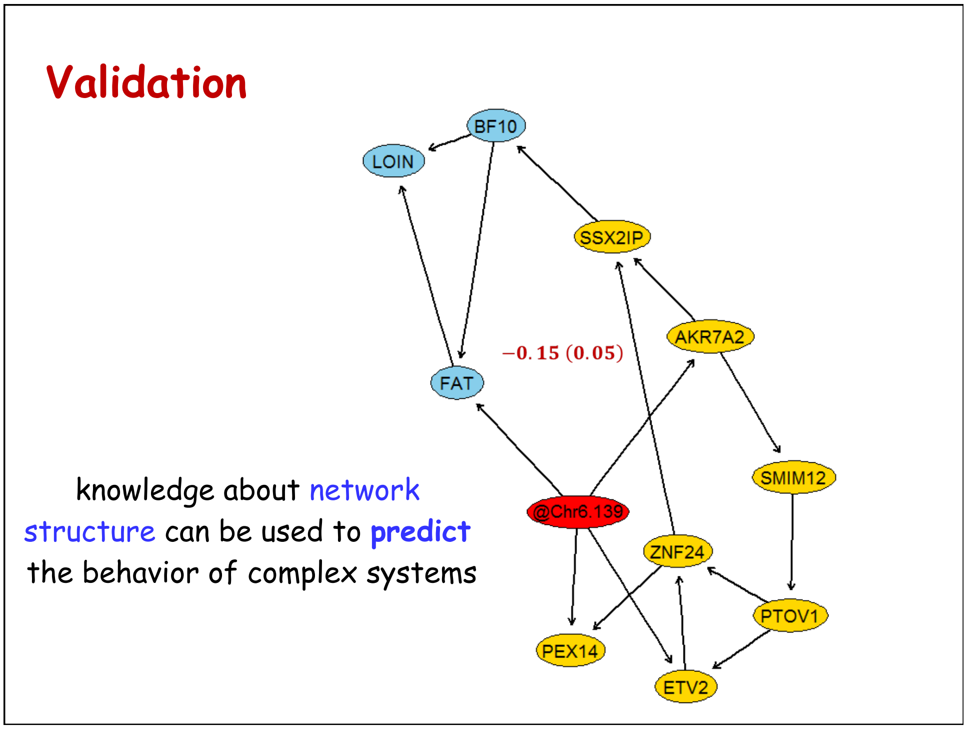
the effects are **mediated by the expression** of several genes



Network Stability

Jackknife resampling

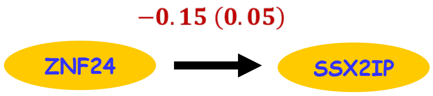
The majority of the links and directions show great stability



Validation

knowledge about network structure can be used to **predict** the behavior of complex systems

Validation



the network predicts that modulation of the expression of ZNF24 should **lead to changes in the expression** of SSX2IP

recent study has **overexpressed/silenced** ZNF24 and then applied microarray assay to identify **target genes**

↑↑ ZNF24 **decreased the expression** of SSX2IP

⊗ ZNF24 resulted in a **overexpression** of SSX2IP

Causal Network 2.0

Integrate **multiple layers of omics data:**
phenotypic, **genotypic**, **transcriptomic**, **metabolomic data**

- whole-genome DNA sequencing data
- RNA-Seq data
- metabolomic data
- multiple phenotypes