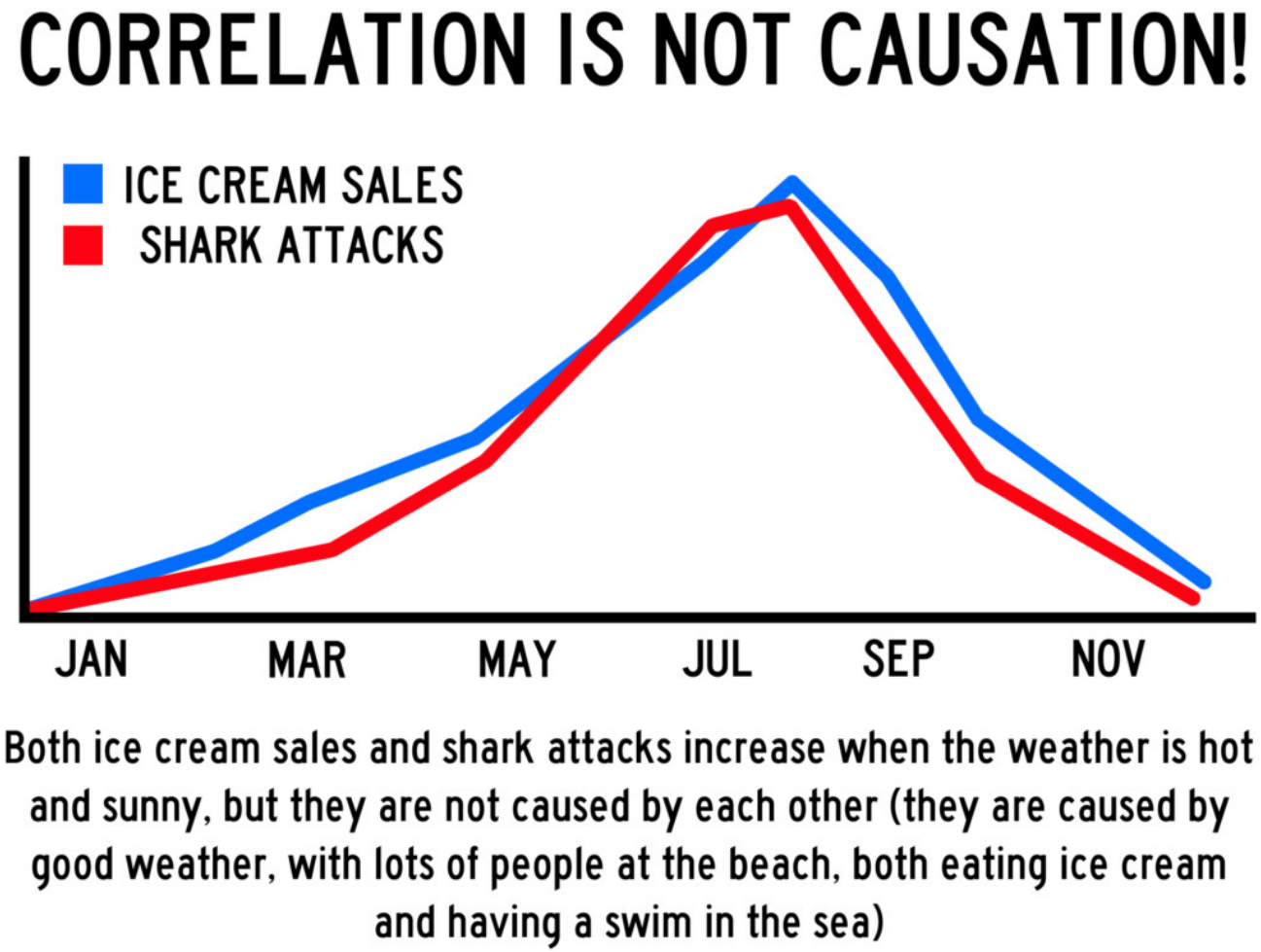
Statistical learning of epistasis, genetic maps and microbial networks via graphical models

Prof. Ernst C. Wit

Institute of Computational Science
Università della Svizzera italiana

Joint work with Pariya Behrouzi, Olaf Schenk and Arnaud Cougoul
16 May 2019

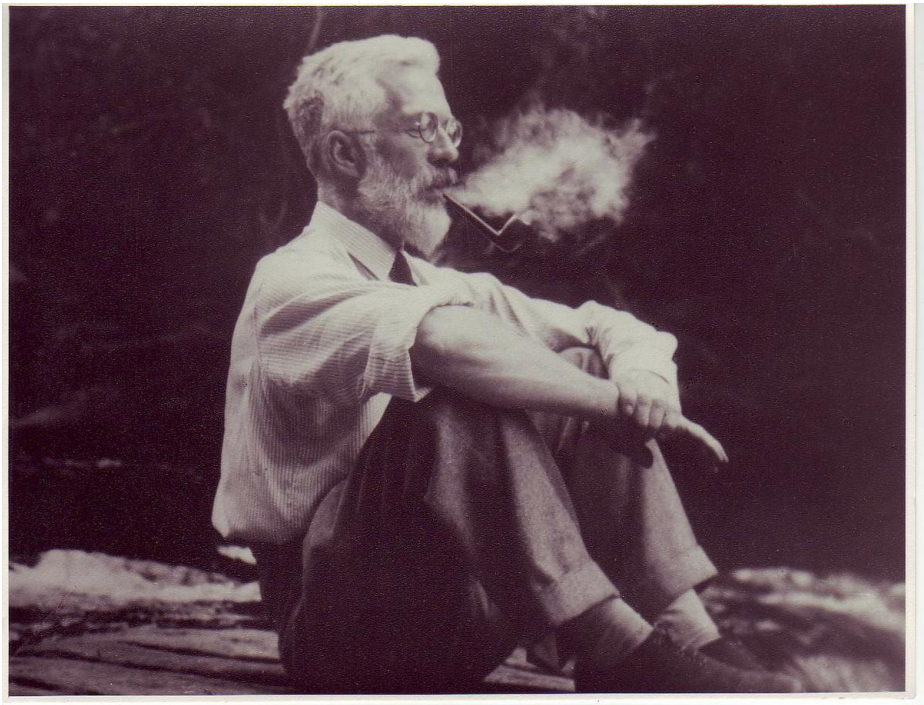
Correlation $\neq$ causation



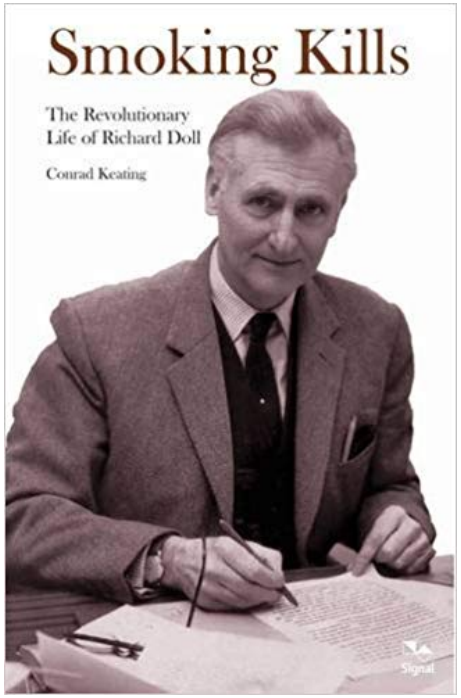
Events are connected by a **common cause**: confounding

R.A. Fisher vs. Richard Doll

R.A. Fisher (geneticist and statistician) was a fervent smoker.



"Smoking and lung cancer are confounded"



"Control for *all* possible confounders"

Sir Richard Doll conducted:

- 1950. Lung cancer study in 20 London hospitals.
- 1954–2001 British Doctor Study to eliminate confounders.

Università della Svizzera italiana

Eliminating confounders

Although in a marginal model

$$\text{lung cancer} = \theta_0 + \theta_1 \text{smoking} + \epsilon,$$

it may seem like

$$\theta_1 > 0 \implies \text{“smoking causes cancer”},$$

but this is the **correlation = causation fallacy**.

Controlling for confounders

Only if in a complete model

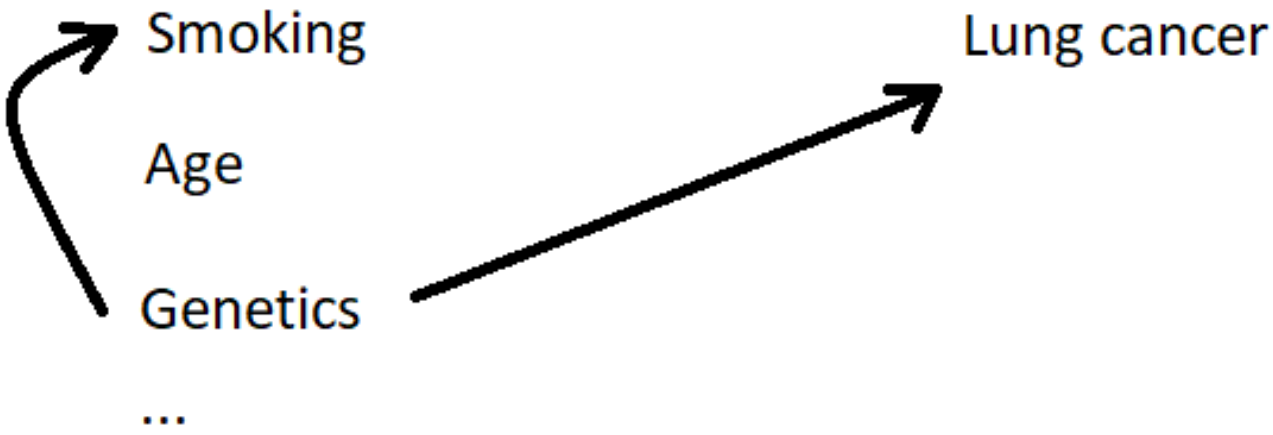
$$\text{lung cancer} = \theta_0 + \theta_1 \text{smoking} + \theta_2 \text{age} + \theta_3 \text{genetics} + \ldots + \epsilon$$

we still have $\theta_1 > 0$, then we can *hope* for causation.

Eliminating confounders (visually)

Instead of only checking for correlation,

we should control for other possible explanations:

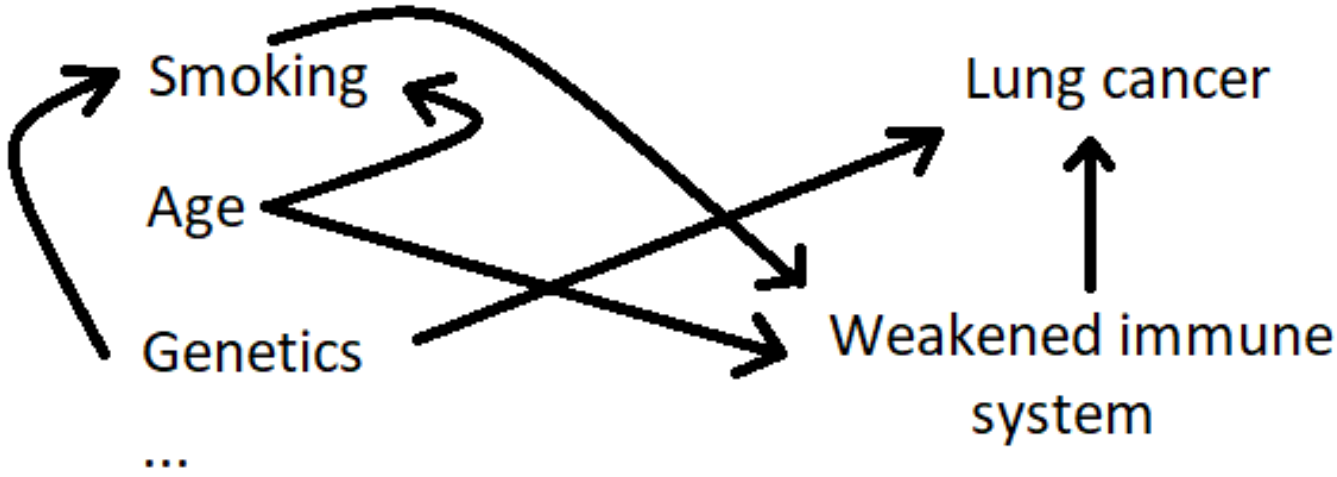


No wonder that the British Doctor Study lasted until 2001!

Ps. Philosophically, it is never finished:
Think Popper's falsification vs verification.

Networks: eliminating confounders systematically

A **direct influence** network can have a causal interpretation,
. . . if all potential confounders are included.



Aim of this talk

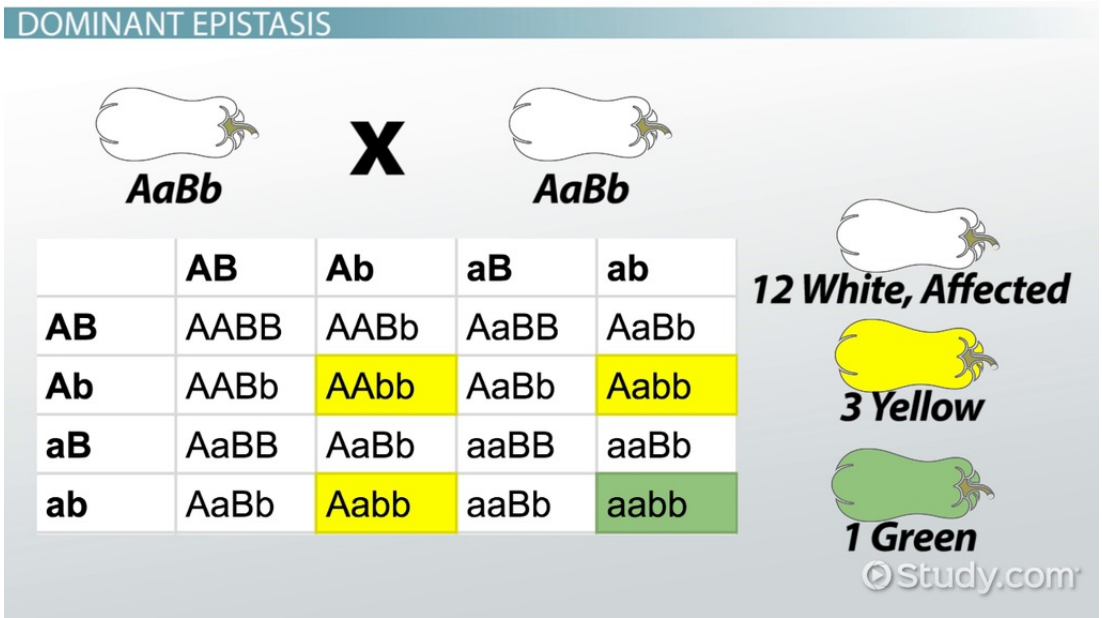
To estimate this (conditional independence) graph from data

1. Epistatic interactions

Epistasis (= "statistical interaction")

Epistasis

when one gene locus masks or modifies a second gene locus phenotype

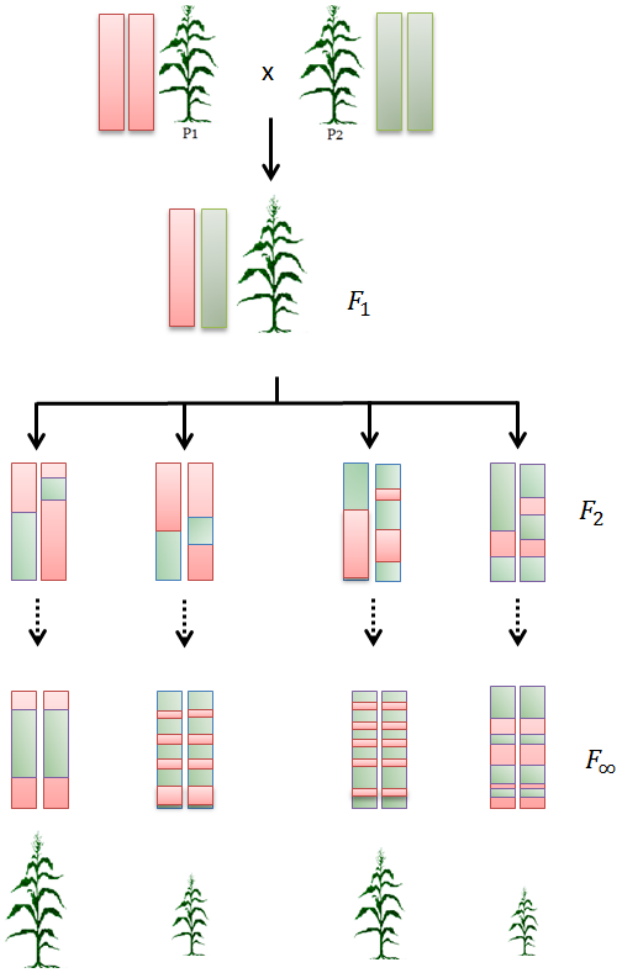


	AB	Ab	aB	ab
AB	AABB	AABb	AaBB	AaBb
Ab	AABb	AAbb	AaBb	Aabb
aB	AaBB	AaBb	aaBB	aaBb
ab	AaBb	Aabb	aaBb	aabb

Phenotype = squash colour

Arabidopsis Recombinant Inbred Line study

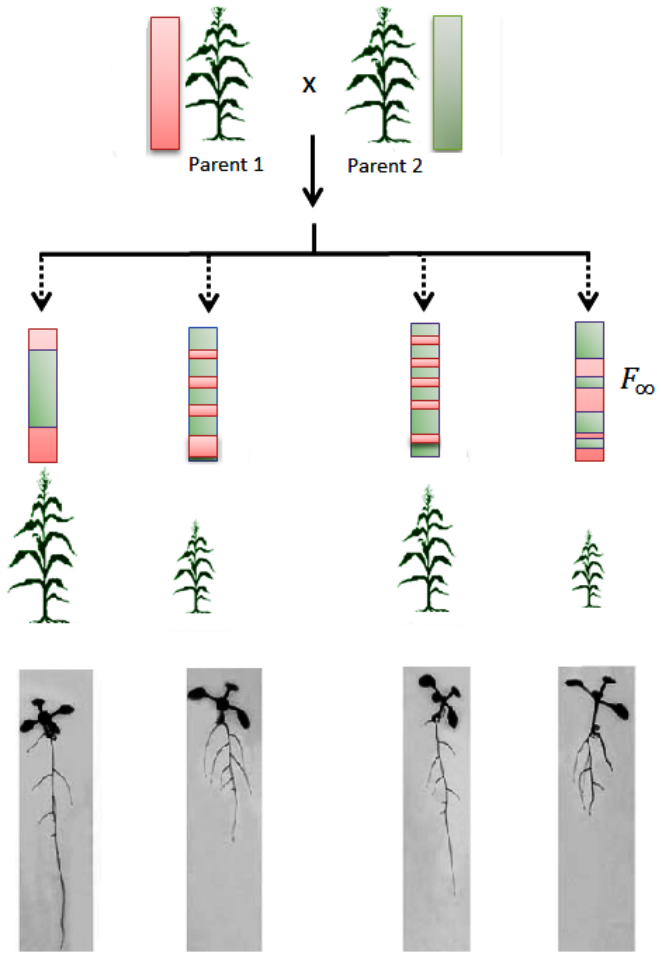
RIL cross made consisting of 367 F_8 generation RIL Arabidopsis plants:



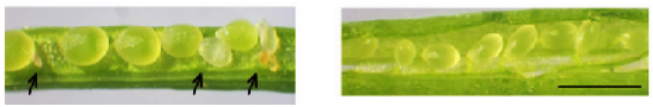
The plants are genotyped at 90 markers.

Università della Svizzera italiana

Epistasis in Arabidopsis



Embryo lethal phenotype



Certain genotype combinations may be functionally incompatible.

Phenotype = survival

Aim of this genotype study

Detection of **high-dimensional epistatic selection** across Arabidopsis genome.

Approach

Extending **graphical model** for **discrete ordinal** data to determine the pattern of **conditional independence** relationships.

RIL genotype data

For each of $i = 1, \ldots, 367$ plants, we obtain genotype at $j = 1, \ldots, 90$ loci:

$$L_j^{(i)} = \text{genotype of marker } j \text{ for plant } i.$$

E.g., in heterozygous diploid population:

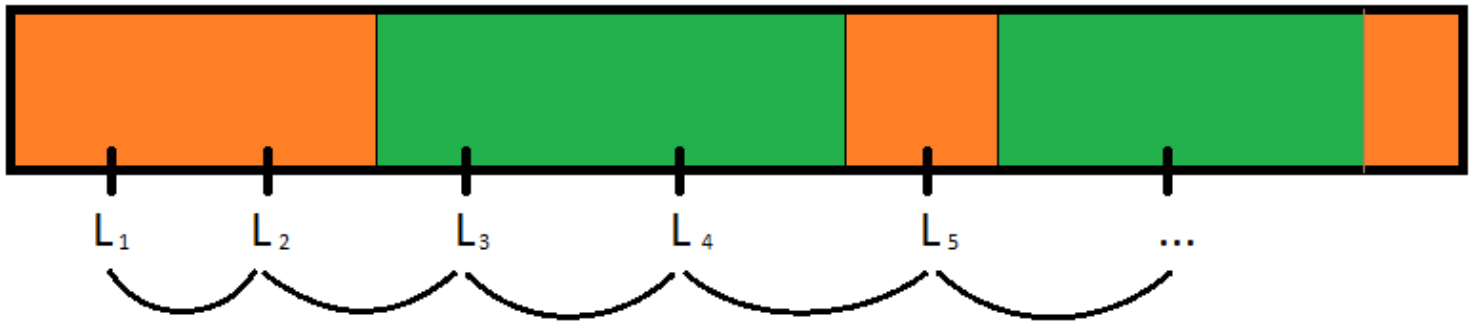
$$L_j = \begin{cases} 0 & \text{homozygous AA from parent 1} \\ 2 & \text{heterozygous AB} \\ 4 & \text{homozygous BB from parent 2} \\ NA & \text{if genotype is missing} \end{cases}$$

PS. Although in RIL genotype is typically homozygous (as in figures), we do have some heterozygous genotypes too.

"Eliminating confounders" idea (1)

Normally, only neighbouring loci are informative about a locus:

$$
\begin{aligned}
L_1 &= \theta_{12} L_2 + \epsilon_1 \\
L_2 &= \theta_{21} L_1 + \theta_{23} L_3 + \epsilon_2 \\
L_3 &= \theta_{32} L_2 + \theta_{34} L_4 + \epsilon_3 \\
\ldots &= \ldots
\end{aligned}
$$



Genetic laws responsible:

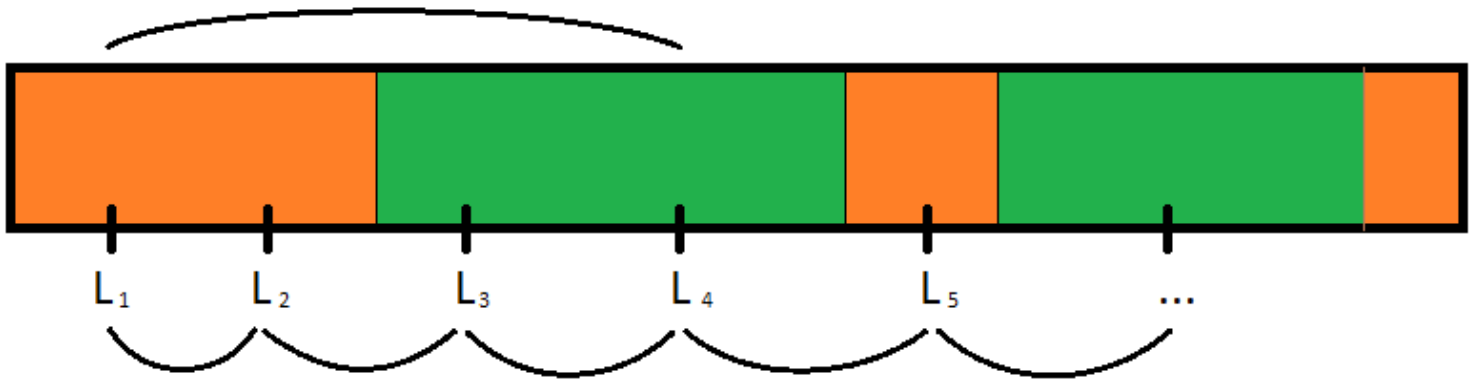
- Genetic linkage
- Independent assortment of chromosomes

Università della Svizzera italiana

“Eliminating confounders” idea (2)

IF non-neighbouring locus is predictive, then possibly sign of **epistasis**:

$$
\begin{aligned}
L_1 &= \theta_{12} L_2 + \theta_{14} L_4 + \epsilon_1 \\
L_2 &= \theta_{21} L_1 + \theta_{23} L_3 + \epsilon_2 \\
L_3 &= \theta_{32} L_2 + \theta_{34} L_4 + \epsilon_3 \\
\ldots &= \ldots
\end{aligned}
$$



Inference idea

The non-zeros in the matrix Θ

$$\Theta = \begin{pmatrix} -- & \theta_{12} & \theta_{13} & \dots \\ \theta_{21} & -- & \theta_{23} & \dots \\ \theta_{31} & \theta_{32} & -- & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

determine the underlying linkage & epistasis structure.

(A bit) more precisely ...

Locus data $L^{(i)} = (L_1^{(i)}, \dots, L_{90}^{(i)})$ can be modelled as

$$L^{(i)} \sim N(\mu, \Theta^{-1}),$$

for $i = 1, \dots, 367$ F_8 Arabidopsis plants.

NOTE:

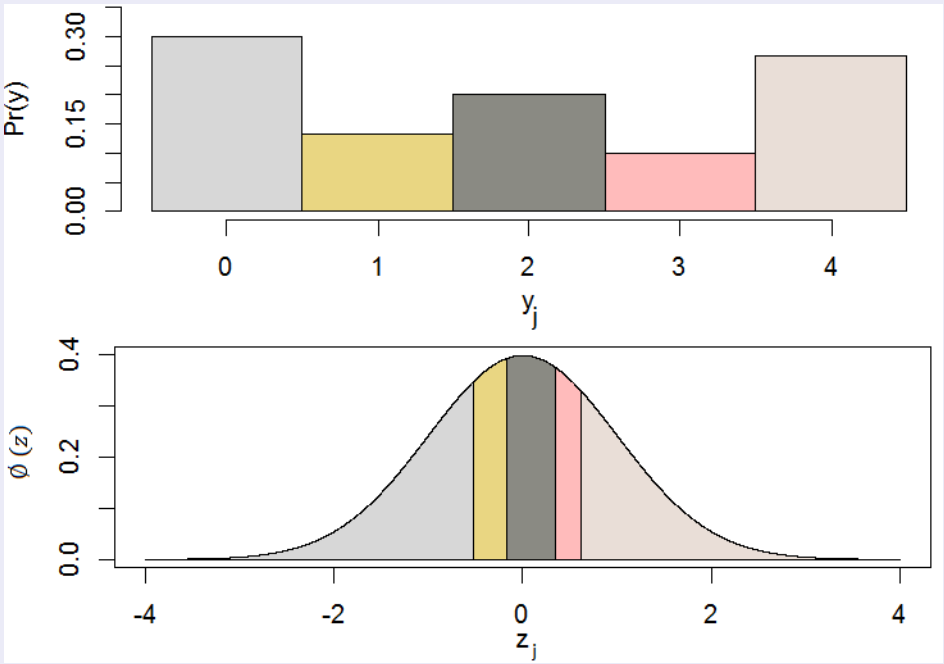
$$\text{Cor}(L_1, L_4 | \text{ rest}) = \frac{-\theta_{14}}{\sqrt{\theta_{11}\theta_{44}}}.$$

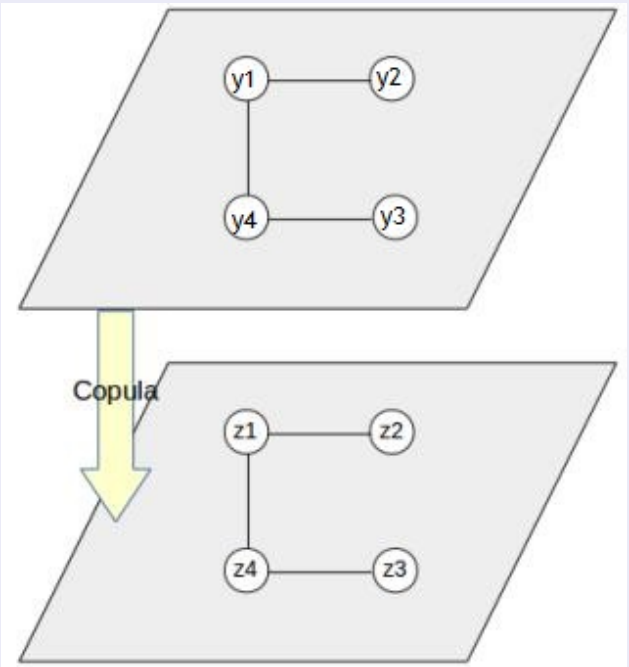
Università della Svizzera italiana

Gaussian Copula

Assume latent variable $Z \sim N(0, \Theta^{-1})$ in the following way:

Relationship between latent and observed variables





- A set of cut-off points
  $-\infty = c_{j,0} < c_{j,1} < c_{j,2} < \ldots < c_{j,k_j} = \infty$
- $l_j^{(i)} = \sum_{l=0}^{k_j-1} l \times \mathbf{1}_{\{c_{j,l} < z_j^{(i)} \leq c_{j,l+1}\}}$
- $L_j = F_j^{-1}(\Phi(Z_j))$

NOTE: Observations $i = 1, \ldots, 367$, Markers $j = 1, \ldots, 90$.

Penalized graphical Gaussian models

To achieve "sparse" graph structure $\widehat{\Theta}$, introduce an ℓ_1 penalty:

$$\widehat{\Theta} = \text{argmax } \log|\mathbf{\Theta}| - \text{Tr}(S\mathbf{\Theta})$$
$$\text{subject to} \sum_{i \neq j} |K_{ij}| \leq \alpha$$

This problem, a.k.a. **Graphical Lasso**, has been considered by

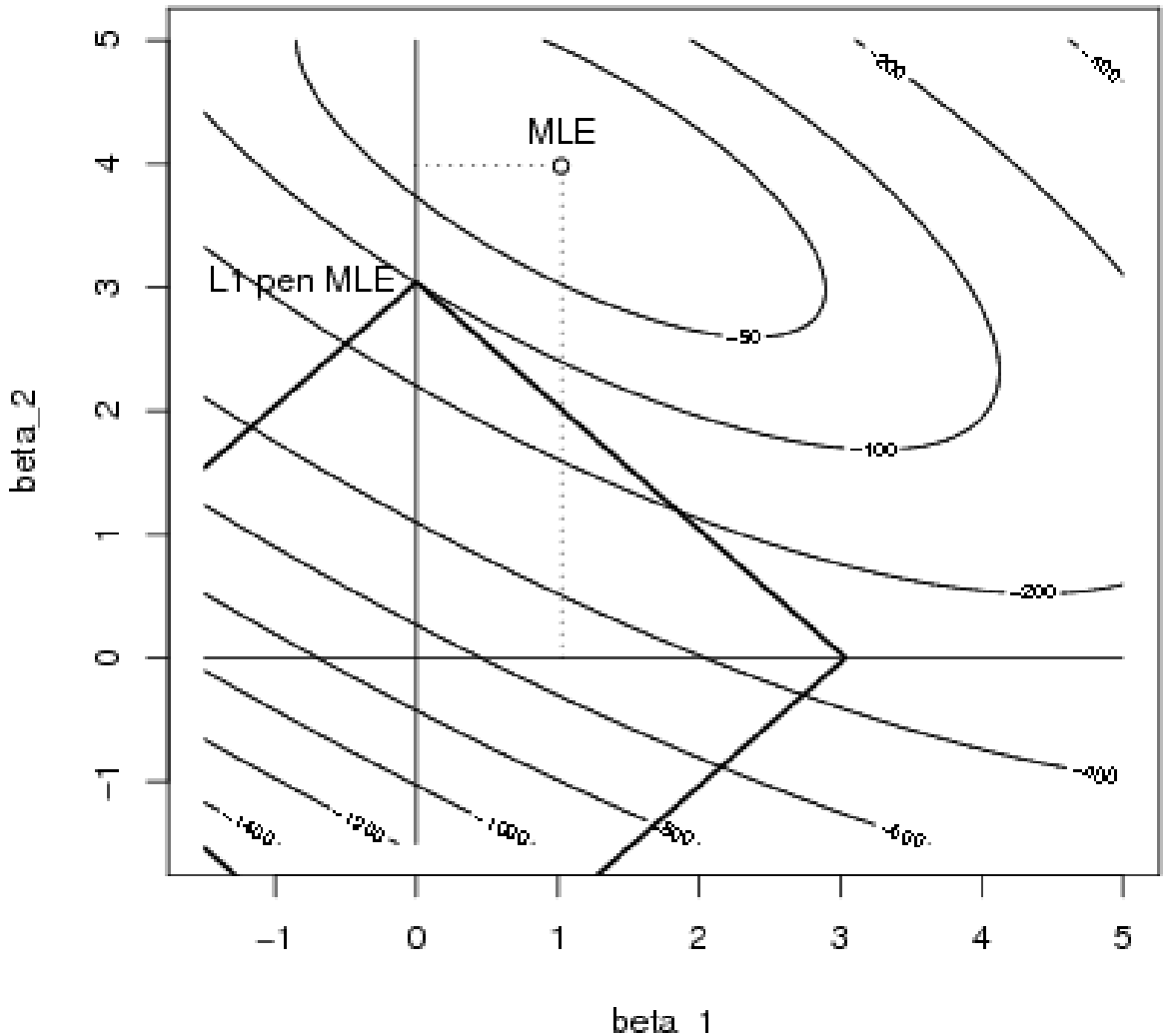
- Meinshausen and Buehlman (2006): implemented in `huge`,
- Banerjee (2007) and
- Friedman et al. (2008): implemented in `glasso`.

Or a Bayesian version:

- Mohammadi and Wit (2015): R-package `BDgraph`.

L_1 penalty = sparsity

Maximize $\ell(\beta) = -(y - X\beta)^t(y - X\beta)$, subject to $||\beta||_1 \leq 3$.



Sparse inference of GCGM

Likelihood:

$$\ell_Y(\Theta) \approx \frac{n}{2} \log \mid \Theta \mid - \frac{1}{2} \sum_{i=1}^{n} \int_{c_{p-1}^{(i)}}^{c_p^{(i)}} \dots \int_{c_1^{(i)}}^{c_1^{(i)}} z^{(i)T} \Theta z^{(i)} dz_1 \dots dz_p$$

Penalized EM algorithm

E-step: Compute $Q_\lambda(\Theta|\Theta^*) = E[\ell^p_{Y,Z}(\Theta)|Y, \Theta^{(m)}]$

$$Q_\lambda(\Theta|\Theta^{(m)}) = -\frac{np}{2} \log 2\pi + \frac{n}{2}\{\log|\Theta| - tr\{\boxed{\frac{1}{n}\sum_{i=1}^{n} E(Z^{(i)}Z^{(i)T}|Y^{(i)}\Theta^{(m)})}\Theta\} - \lambda||\Theta||_1$$

M-step: $\widehat{\Theta} = arg_\Theta \max Q(\Theta|\Theta^*)$ subject to $||\Theta||_1 \leq \lambda$.

della Svizzera italiana

E-step: Approximating conditional expectations fast

$$E(z_k^{(i)} z_l^{(i)^T} \mid y^{(i)}, \hat{\hat{\ }}) \approx \begin{cases} E(z_k^{(i)} \mid y^{(i)}, \hat{\hat{\ }}) E(z_l^{(i)} \mid y^{(i)}, \hat{\hat{\ }}) & \text{if } 1 \leq k \neq l \leq p \\ E(z_k^{(i)^2} \mid y^{(i)}, \hat{\hat{\ }}) & \text{if } k = l \end{cases}$$

$$E(z_j^{(i)} \mid y^{(i)}; \hat{\hat{\ }}, \widehat{C}) \approx \hat{\circ}_{j,-j} \hat{\circ}^{-1}_{-j,-j} E(z_{-j}^{(i)^T} \mid y^{(i)}; \hat{\hat{\ }}, \widehat{C}) + \frac{\phi(\widehat{\delta}^{(i)}_{j,y_j^{(i)}-1}) - \phi(\widehat{\delta}^{(i)}_{j,y_j^{(i)}})}{\Phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}}) - \Phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}-1})} \widetilde{\sigma}_j^{(i)}$$

$$E(z_j^{(i)^2} \mid y^{(i)}; \hat{\hat{\ }}, \widehat{C}) \approx \hat{\circ}_{j,-j} \hat{\circ}^{-1}_{-j,-j} E(z_{-j}^{(i)^T} z_{-j}^{(i)} \mid y^{(i)}; \hat{\hat{\ }}, \widehat{C}) \hat{\circ}^{-1}_{-j,-j} \hat{\circ}^{T}_{j,-j} + \widetilde{\sigma}_j^{(i)^2}$$
$$+ \ldots + \frac{\delta^{(i)}_{j,y_j^{(i)}-1} \phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}-1}) - \widetilde{\delta}^{(i)}_{j,y_j^{(i)}} \phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}})}{\Phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}}) - \Phi(\widetilde{\delta}^{(i)}_{j,y_j^{(i)}-1})} \widetilde{\sigma}_j^{(i)^2}$$

Take-away message: all explicit, so

$$Q_\lambda(\Theta|\Theta^{(m)}) = -\frac{np}{2} \log 2\pi + \frac{n}{2}\{\log|\Theta| - tr\{\widetilde{S}\Theta\} - \lambda||\Theta||_1$$

Università della Svizzera italiana

M-step: Large scale maximization via SQUIC

The orginal QUIC algorithm considers an approximation

$$Q(\Theta + \Delta|\Theta^*) \approx -tr((S - W)\Delta) - \frac{1}{2}tr(W\Delta W\Delta) - \lambda||\Theta + \Delta||_1,$$

for one-dimensional maximization steps $\Delta = c(e_i e_j' + e_j e_i')$.

Bottlenecks of QUIC algorithm:

- dense empirical covariance matrix S is reference for each S_{ij}.
- $\Theta + \Delta$ has to be checked for positive-definiteness.
- $W = \theta^{-1}$ is required.

We now work with Olaf Schenk to extend apply SQUIC method:

1. method is fast
2. on laptop can deal with 100,000 variables (on HPC up to 10 million)

Università della Svizzera italiana

Selection of tuning parameter

$$eBIC_\gamma(\lambda) = -2Q(\widehat{\Theta}_\lambda|\widehat{\Theta}_\lambda^{(m)}) + 2H(\widehat{\Theta}_\lambda|\Theta_\lambda^{(m)}) + df(\widehat{\Theta}_\lambda)$$

where

$$\begin{aligned} H(\widehat{\Theta}_\lambda|\widehat{\Theta}_\lambda^{(m)}) &= E[logL_{Z|z\in\mathcal{D}}(\widehat{\Theta}_\lambda)|z \in \mathcal{D}; \widehat{\Theta}_\lambda^{(m)}] \\ df(\widehat{\Theta}_\lambda) &= (\log n + 4\gamma \log p)d \\ d &= \sum_{1\leq k<l\leq p} I(\widehat{\Theta}_\lambda \neq 0) \ \ and \ \ \gamma \in [0,1]. \end{aligned}$$

Alternatively, a CV based estimator is given by

$$df(\widehat{\Theta}_\lambda) = 2\frac{\sum_{i=1}^n \mathrm{c}[(\widehat{K}_\lambda^{-1} - S_i) \circ I_\lambda]^\top (\widehat{K}_\lambda \otimes \widehat{K}_\lambda)\mathrm{c}[(S - S_i) \circ I_\lambda]}{(n-1)},$$

- $S_k = x_k x_k^T$,
- $I_\lambda = 1 * (\widehat{K}_\lambda \ != 0)$, an indicator matrix.

Università della Svizzera italiana

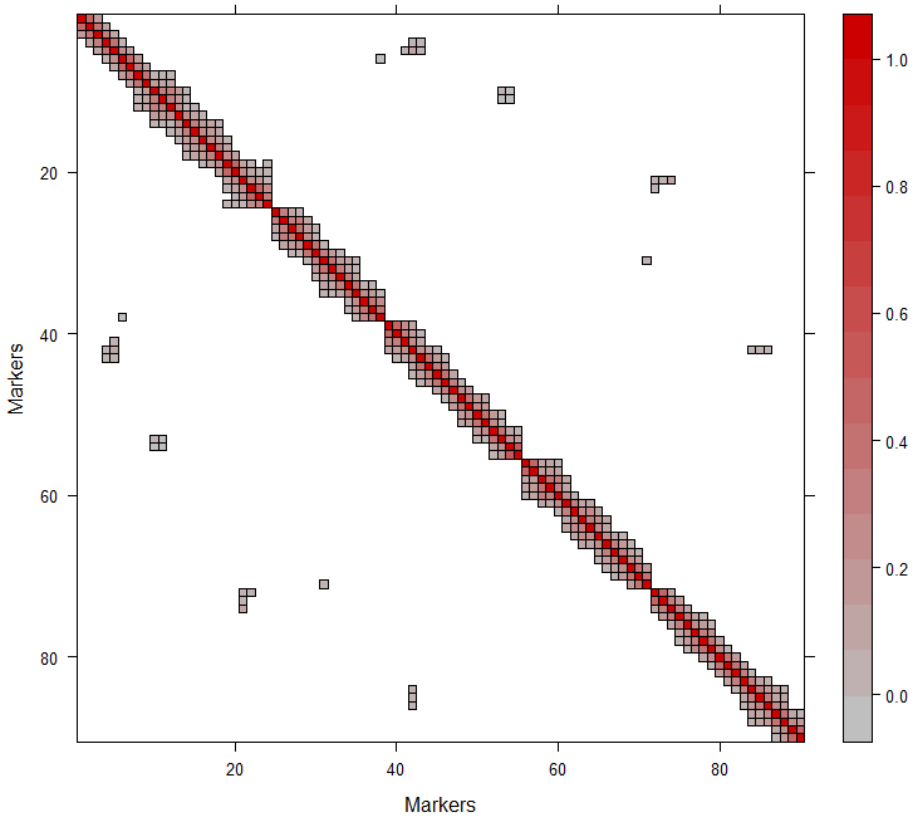
Arabidopsis thaliana experiment

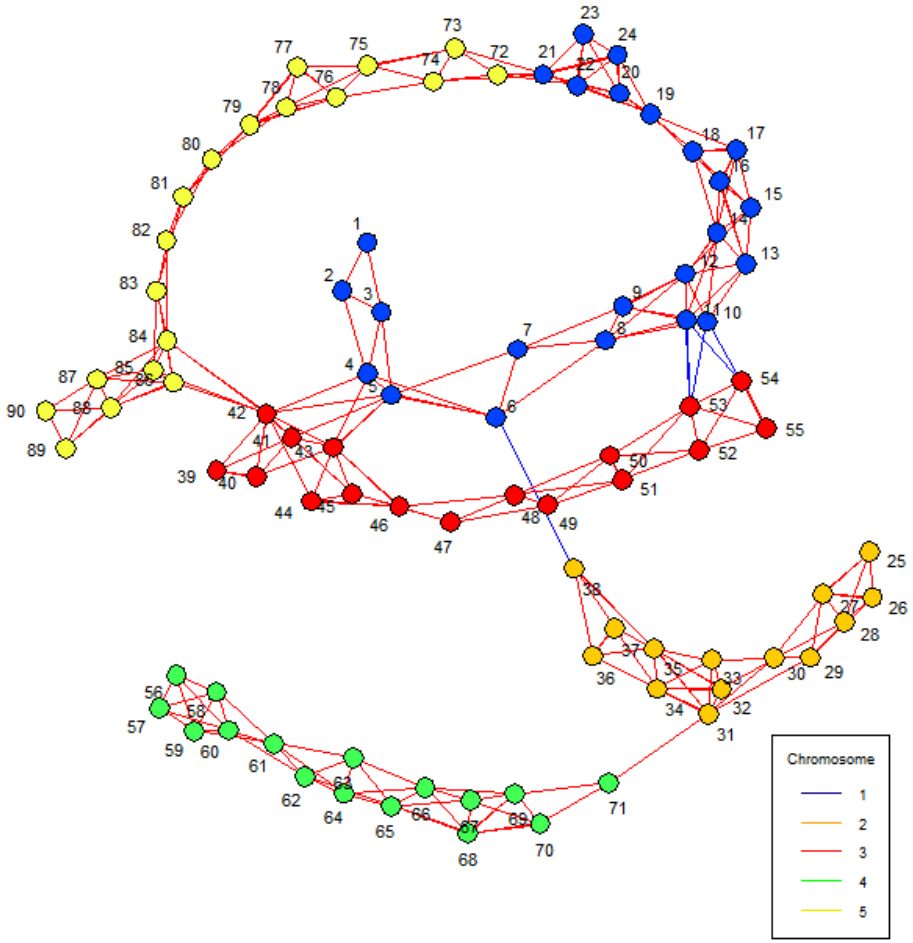
A RIL cross between two **A.thaliana** lines

- Columbia (Col-0) and Cape Verde Island (Cvi-0)
- $\boldsymbol{p = 90}$ SNP markers, $\boldsymbol{n = 367}$ individuals
- Heterozygous population: $\boldsymbol{Y_j^{(i)} \in \{0, 1, 2\}}$
- Contains missing genotypes
- Its genome has 5 chromosomes

Università della Svizzera italiana

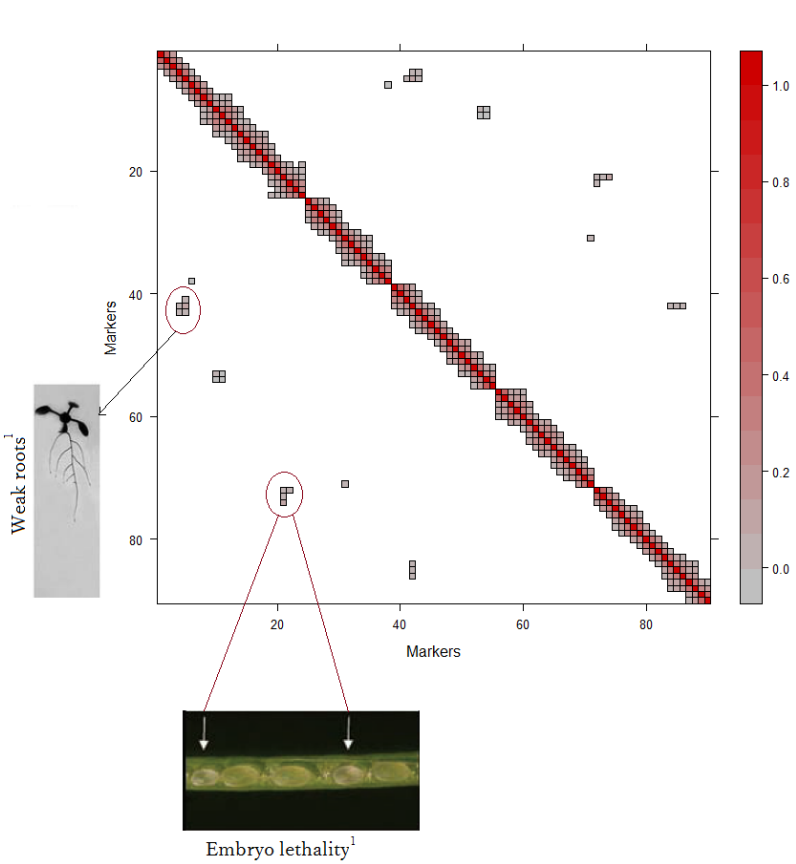
Epistatic selection in *A.thaliana* RIL

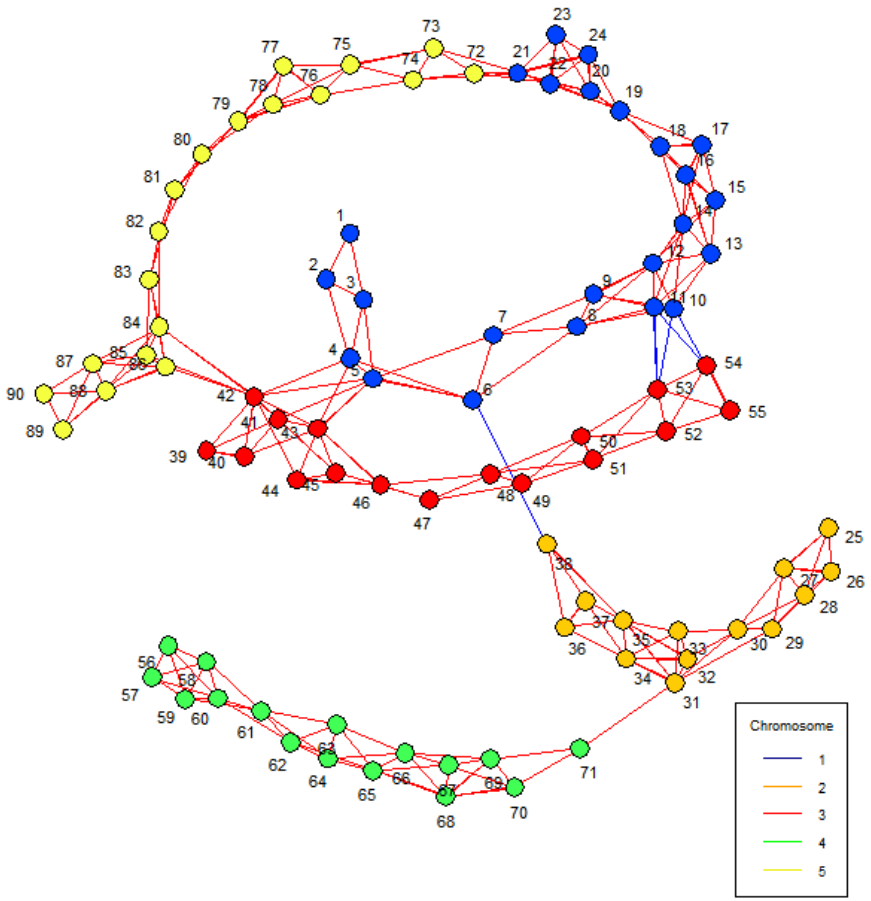




Università della Svizzera italiana

Epistatic selection in *A.thaliana* RIL

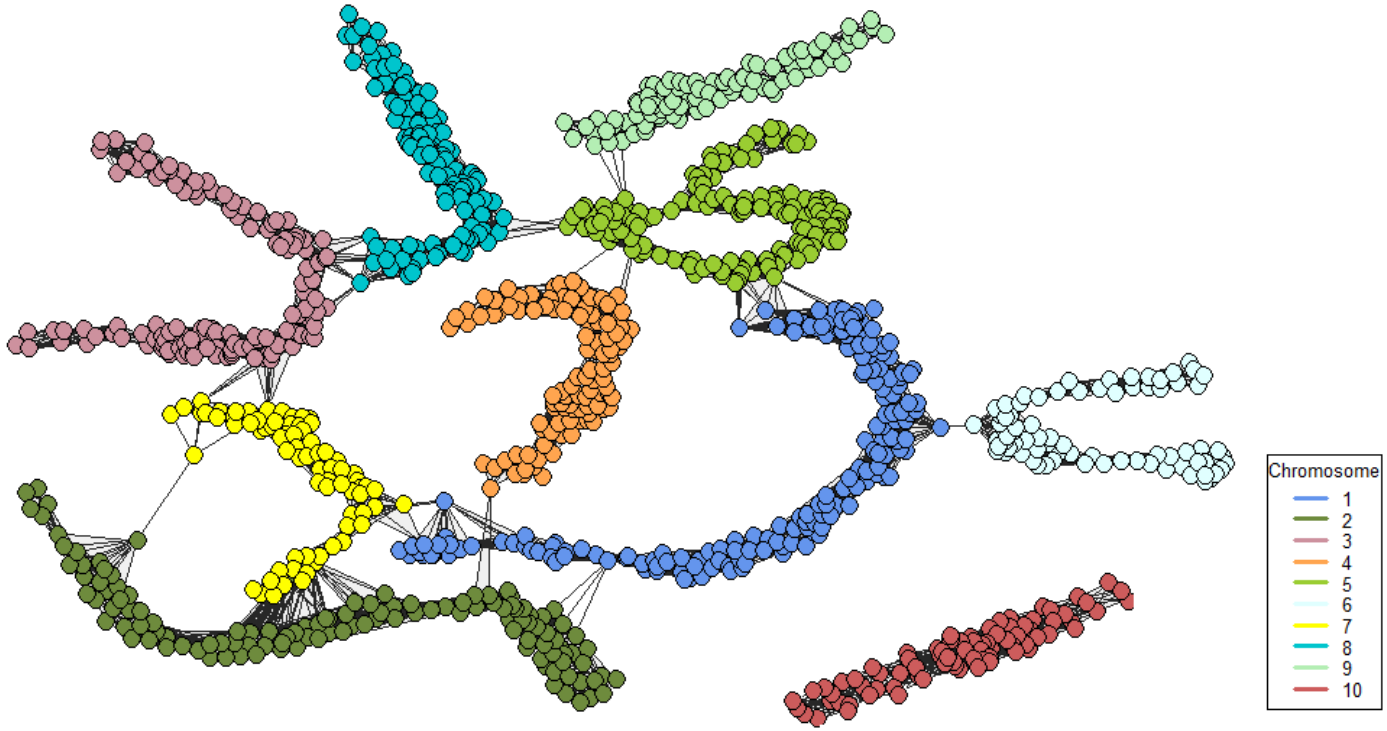




Bikard, David, et al. Science 323.5914 (2009): 623-626.

Genetic inbreeding experiment in Maize

$\boldsymbol{p = 1106}$ SNP markers ; $\boldsymbol{n = 193}$



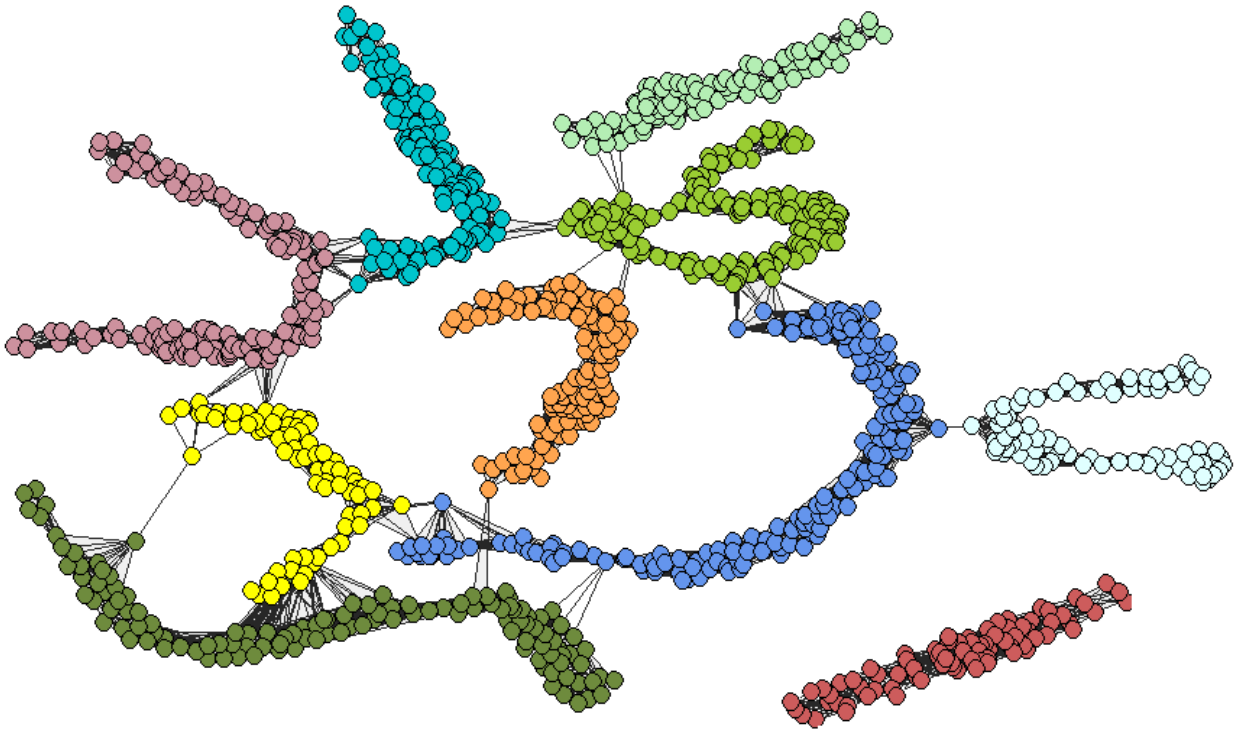
Existence of such trans-chromosomal edges reveals "aberrant" marker-marker associations that are due to epistatic selection.

2. Genetic map construction

From epistasis to genetic maps...

From this picture, it is clear that...

epistasis is a little boat in the ocean of genetic linkage.



Università della Svizzera italiana

... but sampling design is important!

Consider a two parents with genotypes of two loci very close together:

$$AB \times Ab \text{ and } AB \times aB,$$

So all off-spring will have genotype

$$A. \times .B$$

In fact, each of the following will have probability 1/4:

$$Ab \times aB, \quad AB \times aB, \quad Ab \times AB, \quad AB \times AB.$$

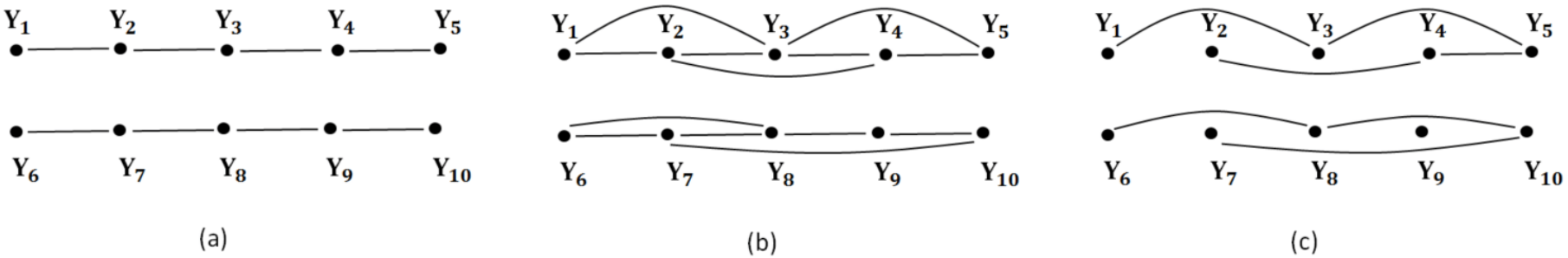
Conclusion

Even though loci are really close together, the off-spring genotype contains NO information of this fact.

3 different population types

Consider

- genotype data Y
- on 10 markers
- of a polyploid species
- from 3 different population types

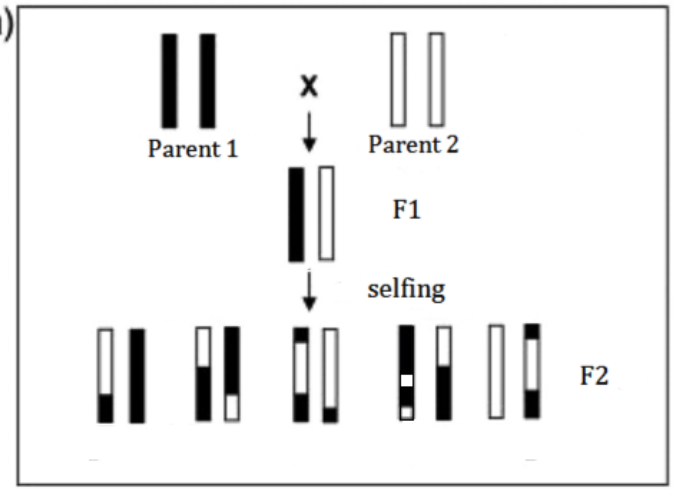


(a) homozygous, (b) inbred, (c) outcrossing (outbred) populations

Università della Svizzera italiana

Mapping algorithm

a)



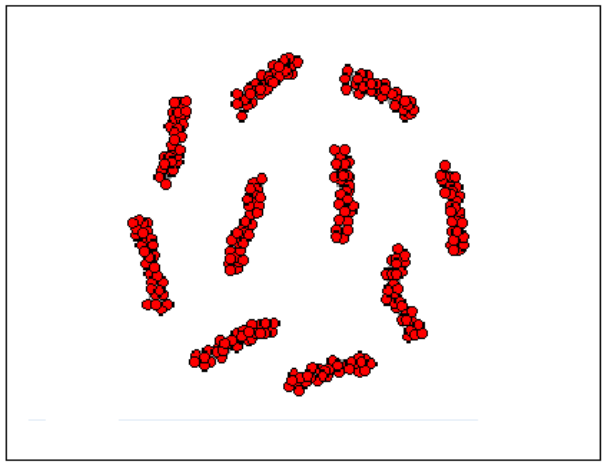
b)

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	...	M500
1	0	0	0	1	1	1	2	2	2	2	1	...	2
2	0	0	0	0	0	-	2	2	2	1	-	...	2
3	2	2	2	2	1	1	0	0	-	0	0	...	-
4	0	-*	0	1	1	1	1	-	0	0	0	...	0
⋮													
200	2	2	2	1	1	1	1	2	2	2	0	...	0

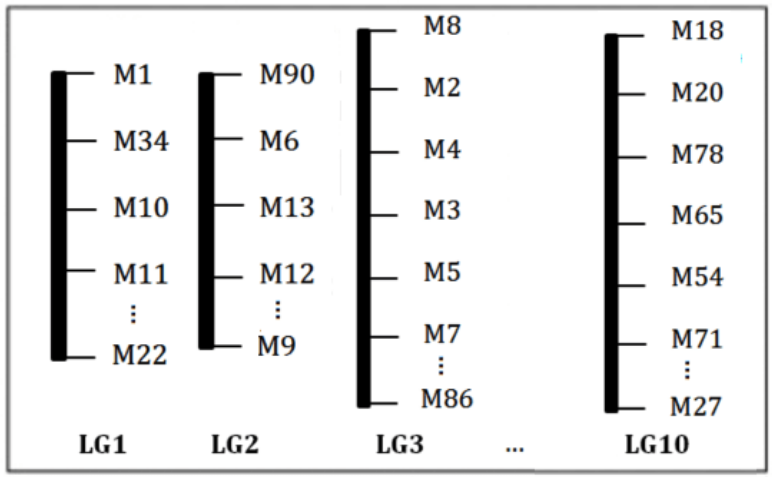
AA= 0, Aa= 1, aa= 2

* Missing genotype

C)

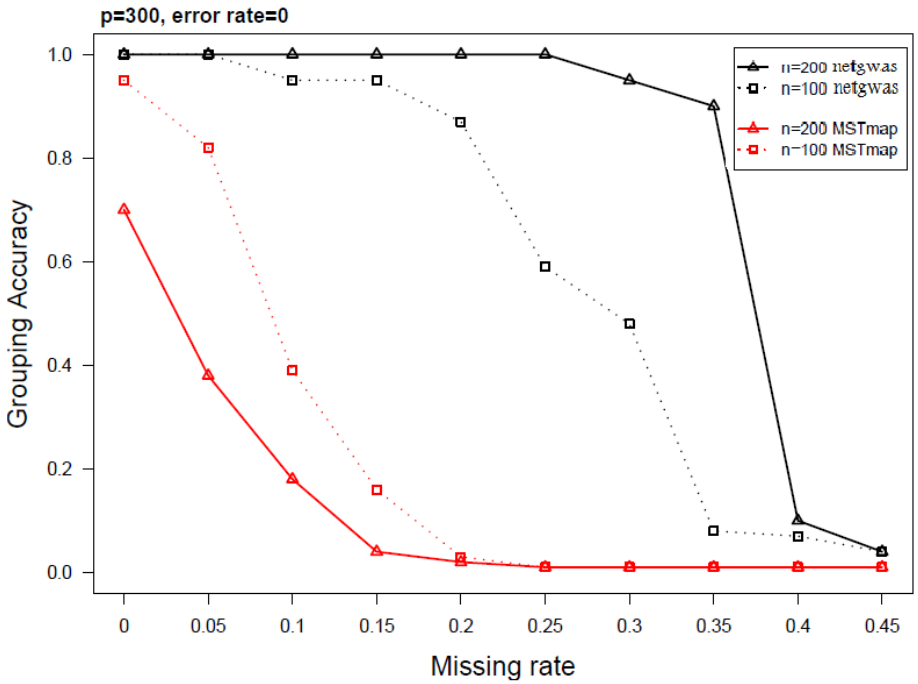


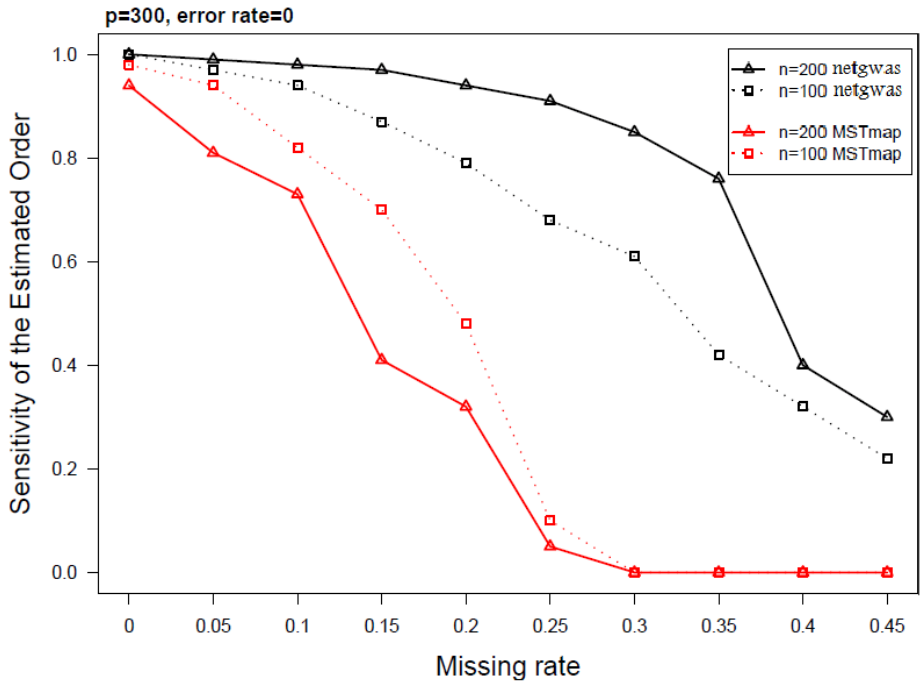
d)



Università della Svizzera italiana

It beats the competition





Università della Svizzera italiana

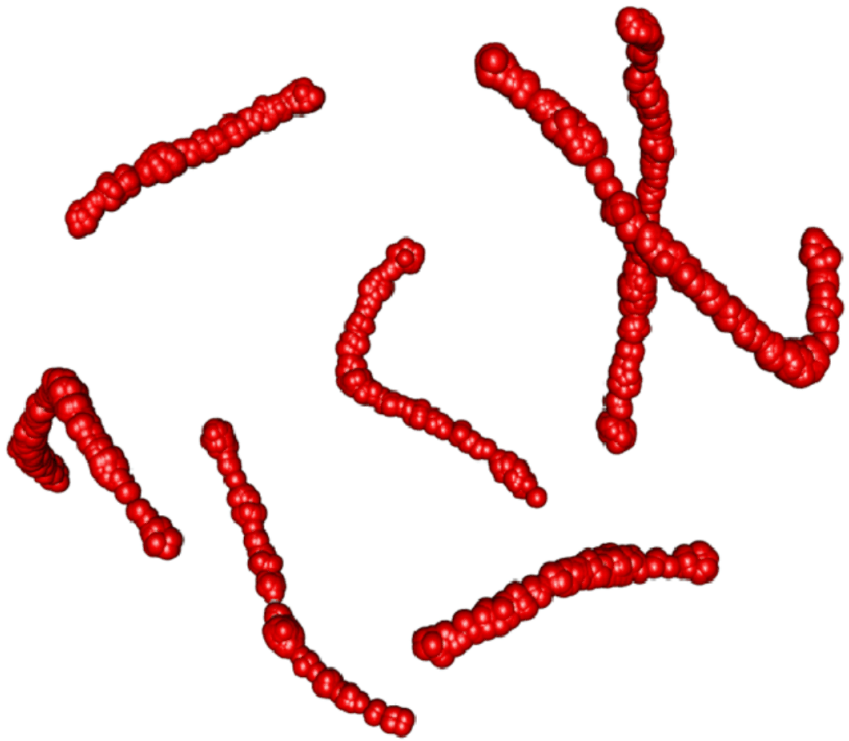
Mapping diploid Barley genome

Estimated number of linkage groups (LGs) for OWB data set

	Estimated # LG	Size of the LGs
netgwas	**7**	**140, 199, 211, 187, 236, 182, 173**
MSTMap	1	1328

Comparison of ordering accuracy between netgwas and MSTMap. In this Table assumed MSTMAP has estimated correctly the number of LGs in the OWB data set.

Linkage Group (LG)	Sensitivity Score	
	netgwas	MSTMap
1	0.86	**0.96**
2	**0.78**	0. 52
3	0.78	**0.92**
4	**0.74**	0.49
5	**0.71**	0.38
6	**0.61**	0.50
7	**0.70**	0.61
Average	**0.74**	0.63

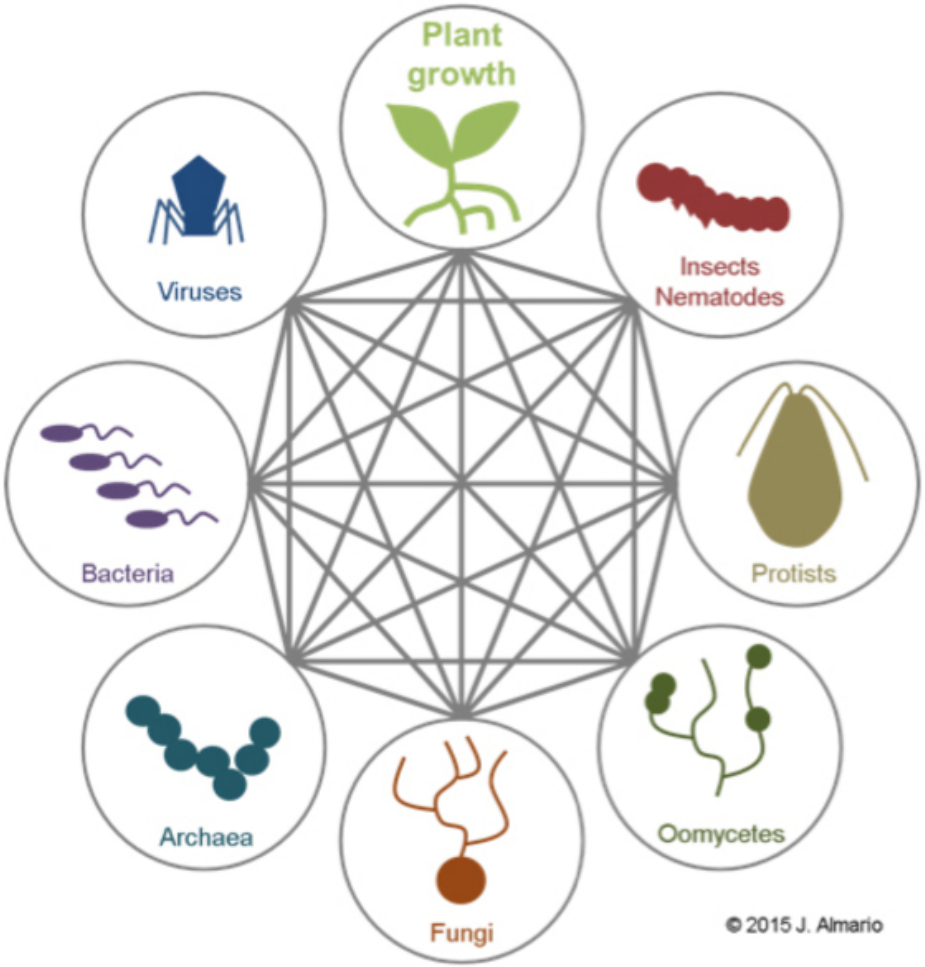


Università della Svizzera italiana

3. Microbial interaction networks

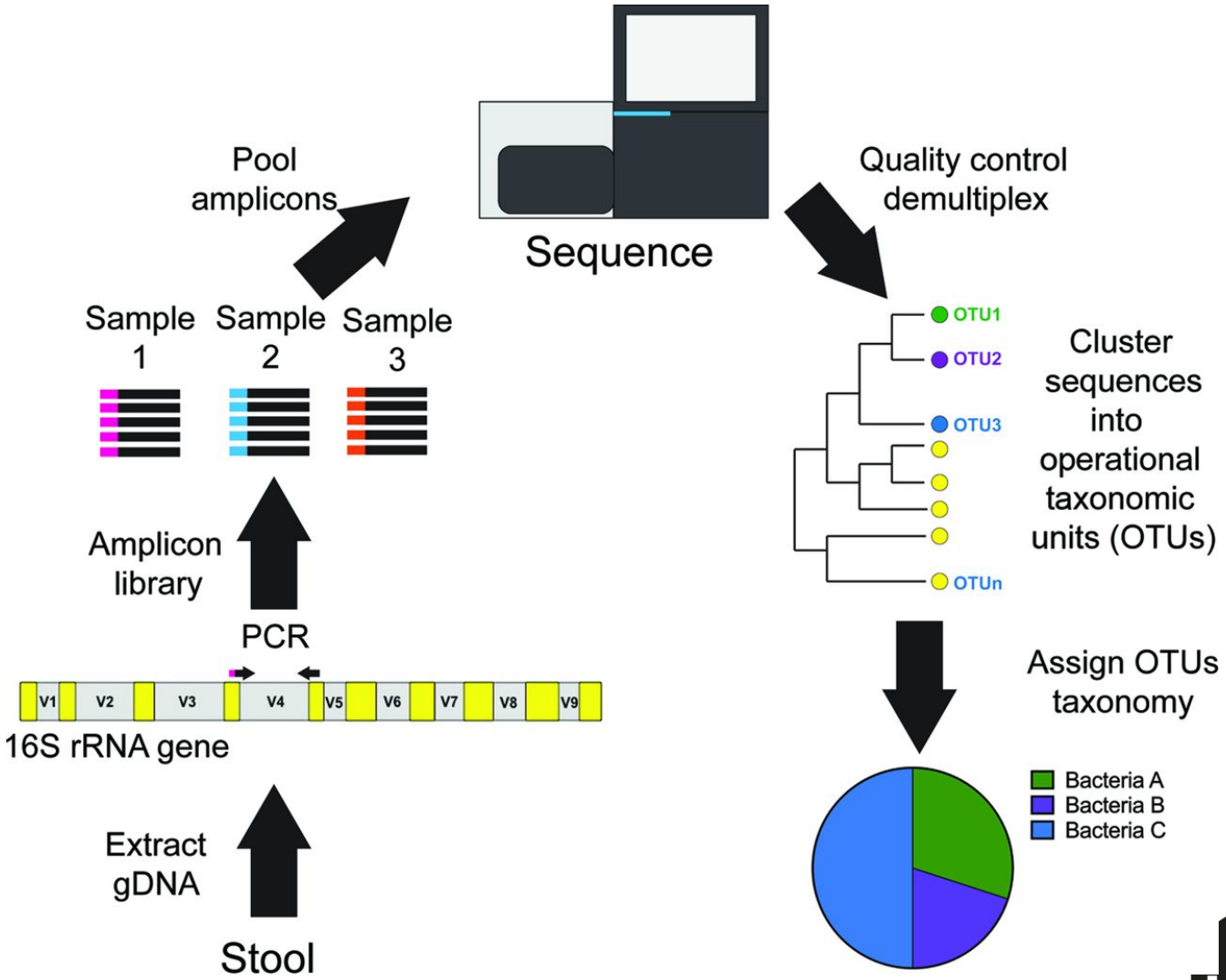
symbiotic relationship

Microorganisms play a central role in many biological processes.



Complex system of interacting species, e.g., involved in plant growth.

Metagenomic data from 16S rRNA sequencing



Università della Svizzera italiana

Human microbiome project

The study collected microbiomes of healthy individual (Methe et al. 2012)

We focus on 306 most prevalent bacteria (OTUs) in 360 stool samples.

Università della Svizzera italiana

Sequence read data

For each stool sample $i = 1, \ldots, 360$, we measure for OTU $j = 1, \ldots, 306$:

$$O_j^{(i)} = \text{read count of OTU } j \text{ for stool sample } i.$$

It is well-known that the read count $O_j^{(i)}$:

- depends on sequencing depth of sample i,
- depends particular peculiarities of OTU j
- has a lot of zero counts
- is overdispersed

We control for all these *nuisance effects*,

$$O_j^{(i)} \sim \text{Zero-inflated negative binomial}$$

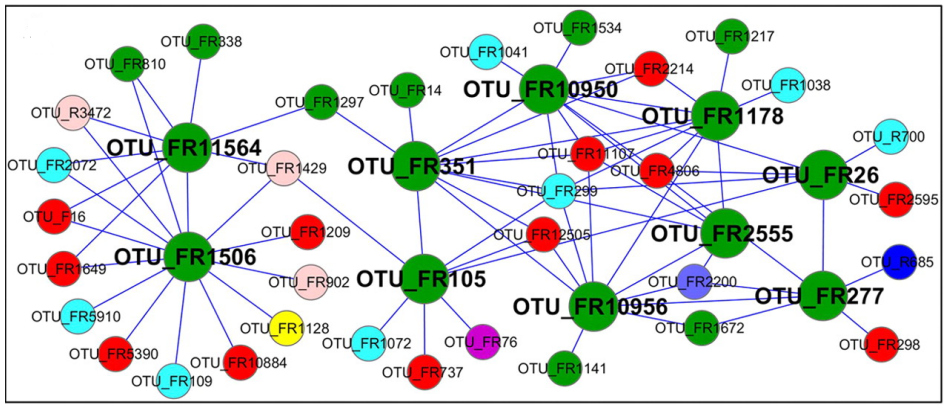
where

$$\log E_\beta O_j^{(i)} = \text{seq. depth}_i + \beta_0 + \beta_j + \ldots$$

Università della Svizzera italiana

Microbial network as Copula Graphical Model

But most importantly, the OTUs depend on each other:

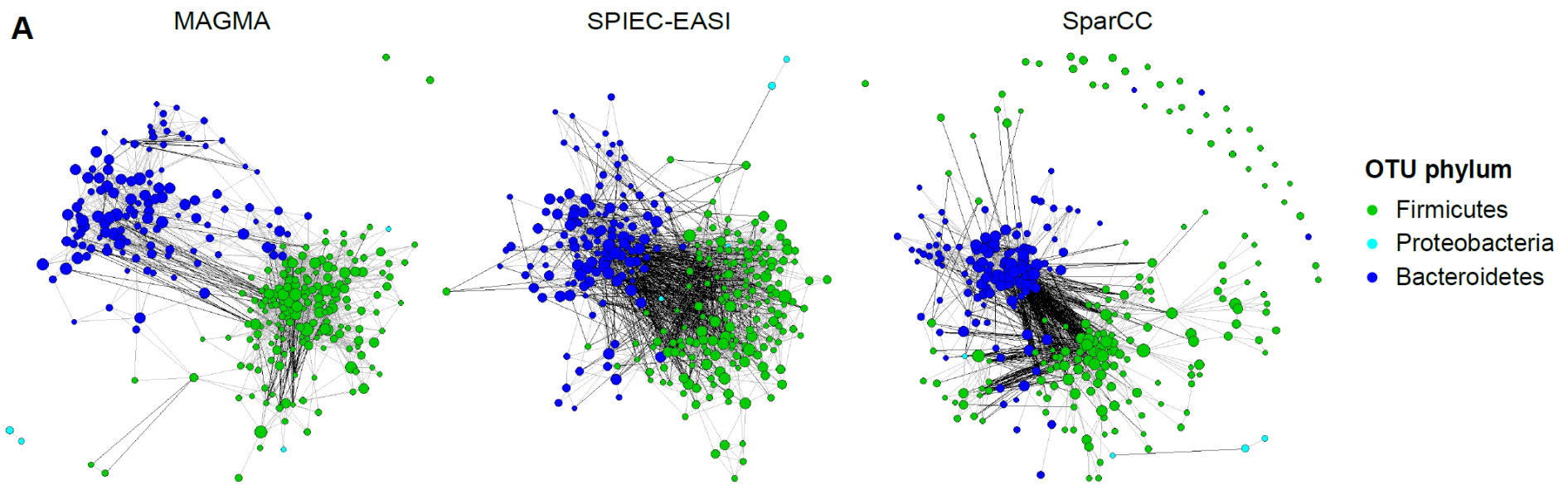


Copula graphical model for interactions:

1. Precision matrix Θ is associated with interaction graph;
2. Latent Gaussians are generated via $Z^{(i)} \sim N(0, \Theta^{-1})$.
3. Observed count $O_j^{(i)}$ is generated by transforming $Z_j^{(i)}$ via ZINB.

Università della Svizzera italiana

MAGMA: inferring microbial interactions



Compared to other methods, our method MAGMA

- Finds fewer spurious links
- Is able to insert real biology in the read count distributions
- Accounts for variability in sequencing depth between samples
- resulting in a cleaner interpretation of the results

Università della Svizzera italiana

Wrapping up

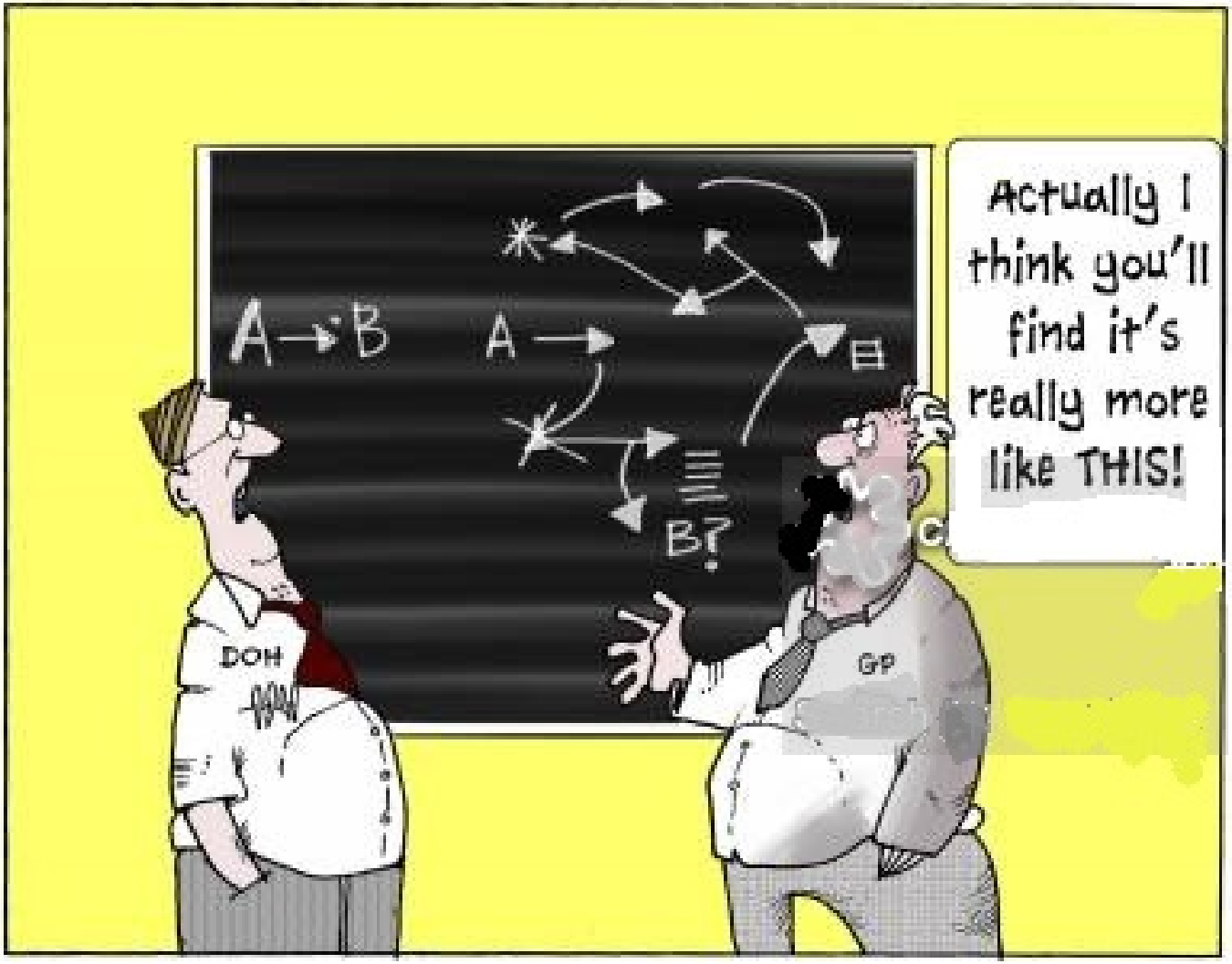
Conclusions

- Confounding is the real enemy of causality.
- Networks account for confounding
  - Detection of **epistatic selection**
  - Construct **genetic map** in in any polyploid species.
  - Reconstruct **microbial interaction networks** from sequence read counts.

Software

netgwas: Our multi-core R package is available on CRAN

rMAGMA: Our microbial network R package is available on github.


A→B
A →
B?
DOH
GP
Actually I think you'll find it's really more like THIS!