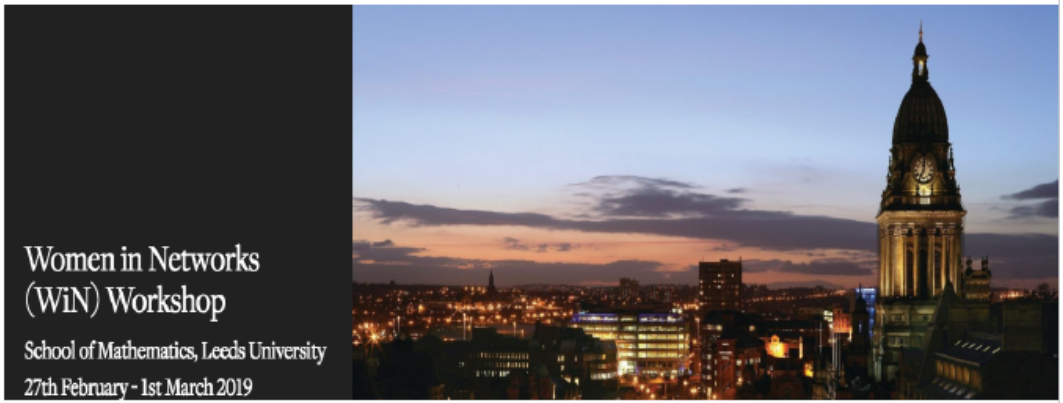
Construction of High-resolution Linkage Maps Using Discrete Graphical Models

Pariya Behrouzi

Mathematical and Statistical Method, Wageningen University, Netherlands



cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

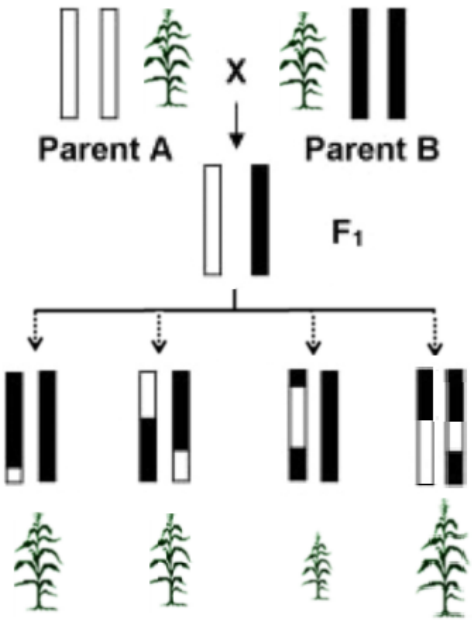
Motivation

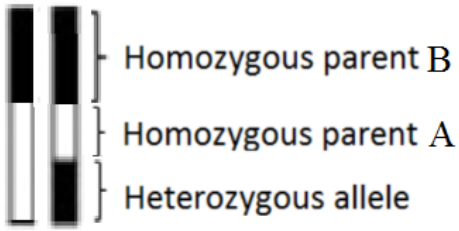
what is linkage map?

Linkage map is order of genetic markers on a chromosome, which contains following info

- **Number of chromosomes** of an species
- Number of **markers inside each chromosome** of the species
- **Order of markers** within each chromosome

Data:





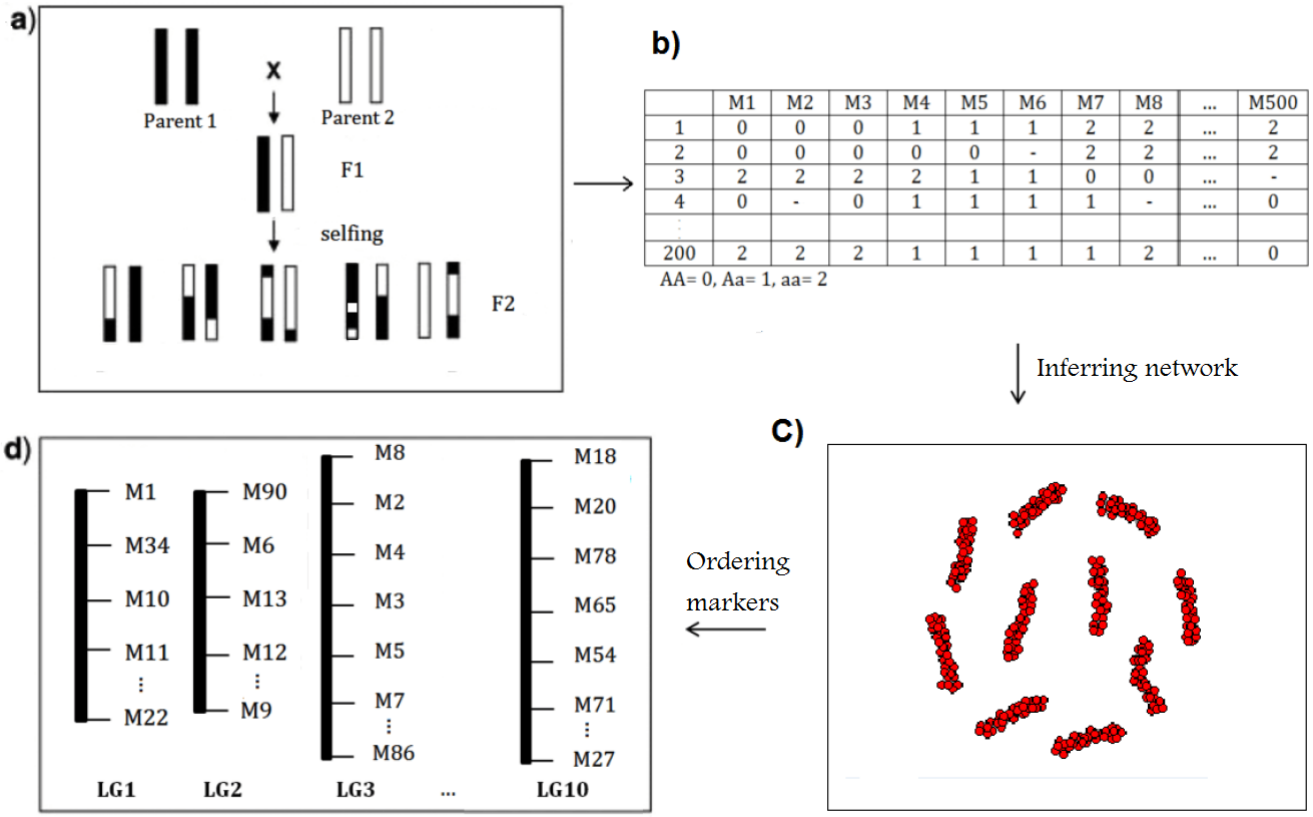
Y_j: observed number of B allele at location j,

$$Y_j = \sum_{k=1}^{q} X_{jk}$$

e.g. for diploids: q=2

Y_j	$X_{j.}$
0	AA
1	AB, BA
2	BB

Motivation



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

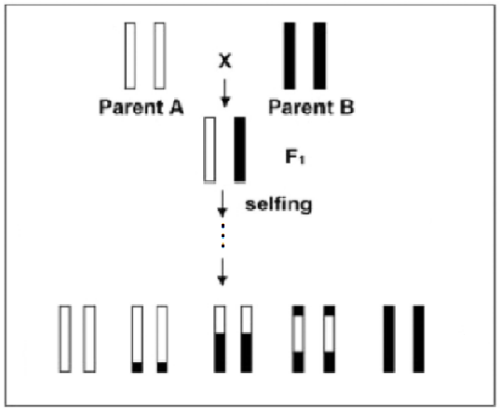
Aim

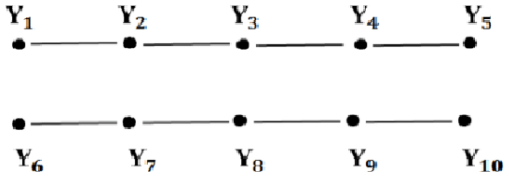
Construct **linkage maps** for species with any copies of chromosome.

Approach

Extending **graphical model** for **ordinal variables** to determine pattern of **conditional independence** among markers.

Meiosis and Markov dependence





Meiosis and Markov dependence

Assume a sequence of ordered markers $X_1, X_2, \ldots, X_{10}$

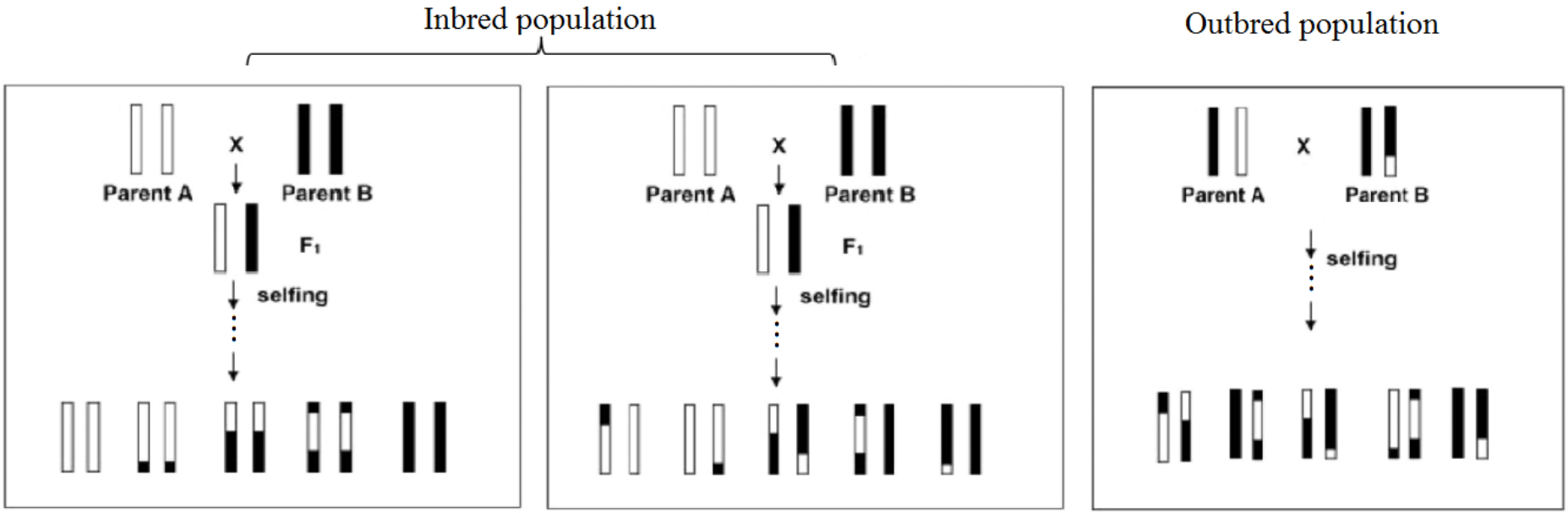
$$Pr(Y_3 \mid Y_1, Y_2, \ldots, Y_{10}) = Pr(Y_3 \mid Y_2, Y_4)$$

$$Y_3 \perp\!\!\!\perp (Y_1, Y_5, \ldots, Y_{10}) \mid (Y_2, Y_4)$$

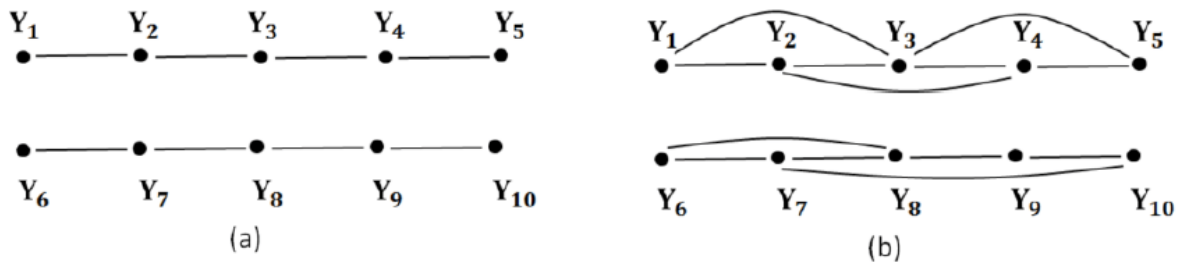
Discrete graphical model

Joint distribution $P(Y)$ can be factorized as: $P(Y) = \prod_{c=1}^{C} \prod_{j=1}^{p} f_{j,j+1}^{(c)}(Y_j^{(c)}, Y_{j+1}^{(c)})$

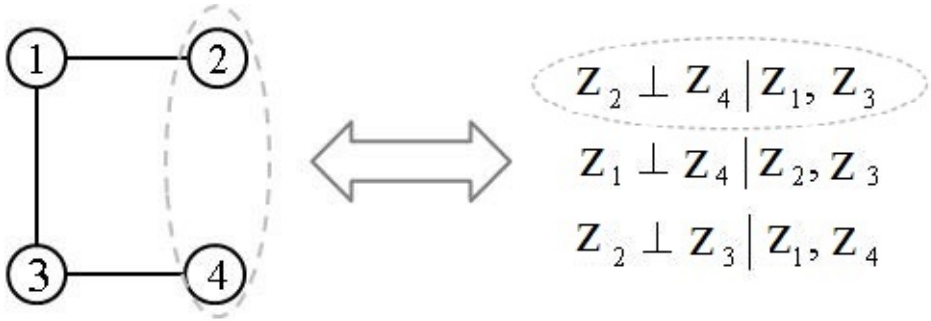
Complications in conditional dependence relationships



Conditional dependence pattern beten neighboring markers:



Gaussian graphical models

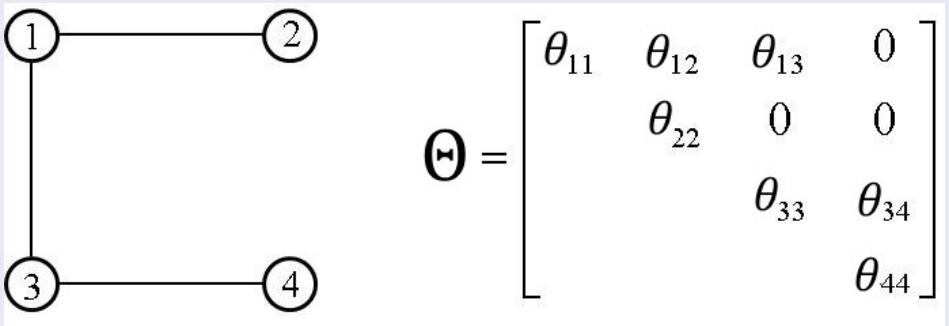


Graph G=(V,E) as

$Z^{(1)}, \ldots, Z^{(n)} \sim \mathcal{N}_p(0, \Sigma), \quad \Theta = \Sigma^{-1}$ is positive definite based on G

Relationship graph, conditional independence and Θ

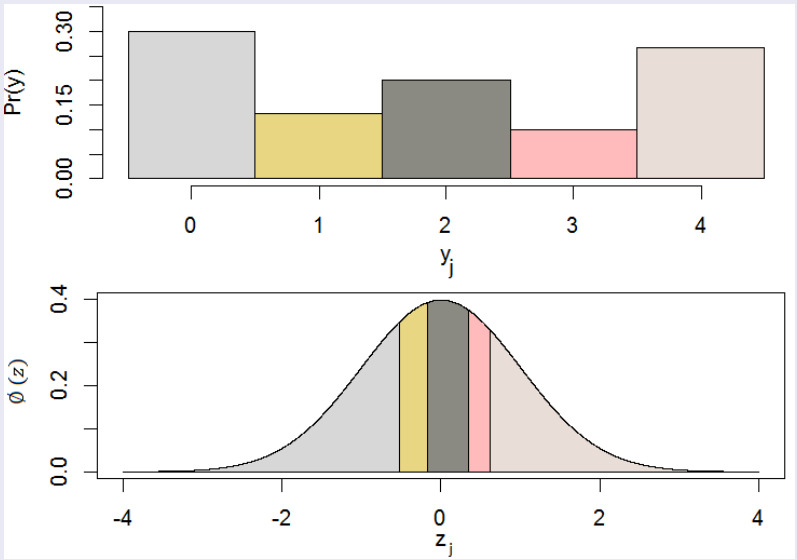
$$Z_i \perp Z_j \mid Z_{V \setminus \{i,j\}} \quad \Leftrightarrow \quad \theta_{ij} = 0$$

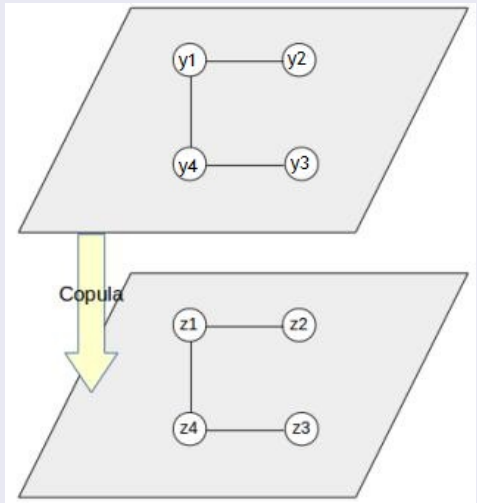


Gaussian Copula

Assume latent variable $Z_j \sim N_p(0, \Theta^{-1})$, $\Theta^{-1} = \Sigma$, $\Sigma_{jj} = 1$ underlying Y_j, where data are $\{y_j^{(i)} | \quad i = 1, \ldots, n, \quad j = 1, \ldots, p\}$

Relationship between latent and observed variables





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

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

$$Q_\lambda(\Theta|\Theta^{(m)}) = -\frac{np}{2}\log 2\pi + \frac{n}{2}\{\log|\widehat{\Theta}_\lambda| - tr\{\boxed{\frac{1}{n}\sum_{i=1}^{n} E(Z^{(i)}Z^{(i)T}|Y^{(i)}\Theta^{(m)})}\}\widehat{\Theta}_\lambda\} - \sum_{j\neq j'}^{p}\omega_{jj'}|\theta|_{jj'}$$

M-step: $\widehat{\Theta}_\lambda = arg_\Theta \max Q_\lambda(\Theta|\Theta^{(m)})$

Estimating conditional expectation

Estimating conditional expectation

1. Gibbs sampling
2. Approximation estimation:

$$E(z_j^{(i)} z_{j'}^{(i)^T} \mid y^{(i)}, \widehat{\Theta}_\lambda) \approx \begin{cases} E(z_j^{(i)} \mid y^{(i)}, \widehat{\Theta}_\lambda) E(z_{j'}^{(i)} \mid y^{(i)}, \widehat{\Theta}_\lambda) & \text{if } 1 \leq j \neq j' \leq p \\ E(z_j^{(i)^2} \mid y^{(i)}, \widehat{\Theta}_\lambda) & \text{if } j = j' \end{cases}$$

Lemma (Johnson et.al (1995))

Let $Z \sim \mathcal{N}(\mu_0, \sigma_0^2)$ *such that* $\delta_1 = (c_1 - \mu_0)/\sigma_0$ *and* $\delta_2 = (c_2 - \mu_0)/\sigma_0$ *for* $c_1 < c_2$

$$E(z_j^{(i)} | c_1 \leq z_j^{(i)} \leq c_2; \widehat{\Theta}_\lambda) = \mu_0 + \frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)} \sigma_0 \tag{1}$$

$$E(z_j^{(i)^2} | c_1 \leq z_j^{(i)} \leq c_2; \widehat{\Theta}_\lambda) = \mu_0^2 + \sigma_0^2 + 2 \frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)} \mu_0 \sigma_0 + \frac{\delta_1 \phi(\delta_1) - \delta_2 \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)} \sigma_0^2 \tag{2}$$

Selection of tunning parameter

At **EM convergence** for a given value of λ

- $\ell_Y(\widehat{\Theta}_\lambda) = Q(\widehat{\Theta}_\lambda|\widehat{\Theta}_\lambda^{(m)}) - H(\widehat{\Theta}_\lambda|\widehat{\Theta}_\lambda^{(m)})$
- $H(\widehat{\Theta}_\lambda|\widehat{\Theta}_\lambda^{(m)}) = E[logL_{Z|z\in\mathcal{D}}(\widehat{\Theta}_\lambda)|z \in \mathcal{D}; \widehat{\Theta}_\lambda^{(m)}]$

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

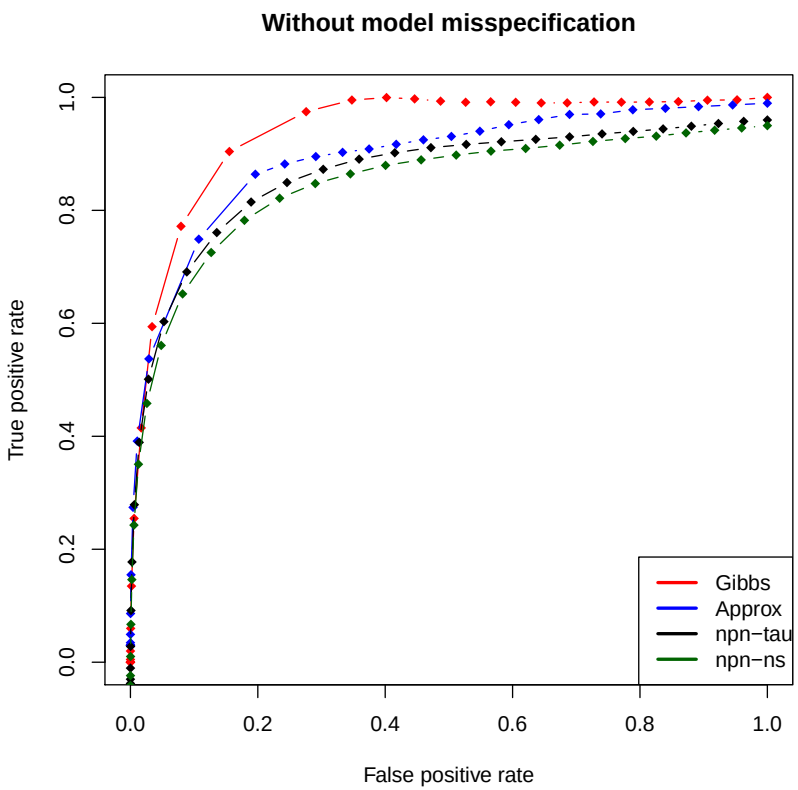
where

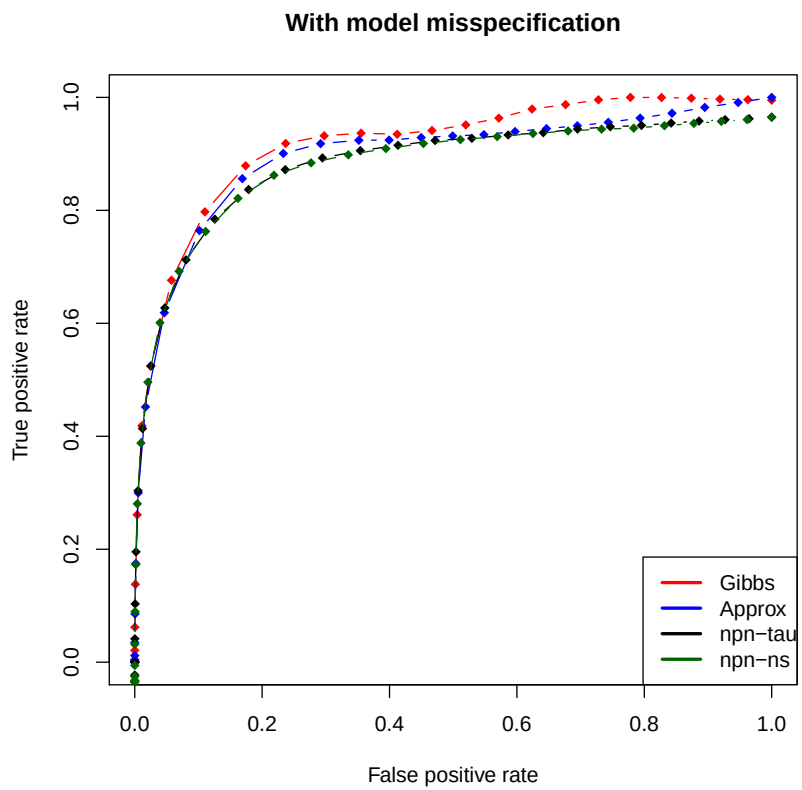
$$df(\widehat{\Theta}_\lambda) = (\log n + 4\gamma \log p)d$$

$$d = \sum_{1\leq k<l\leq p} I(\widehat{\Theta}_\lambda \neq 0)$$

ROC curve

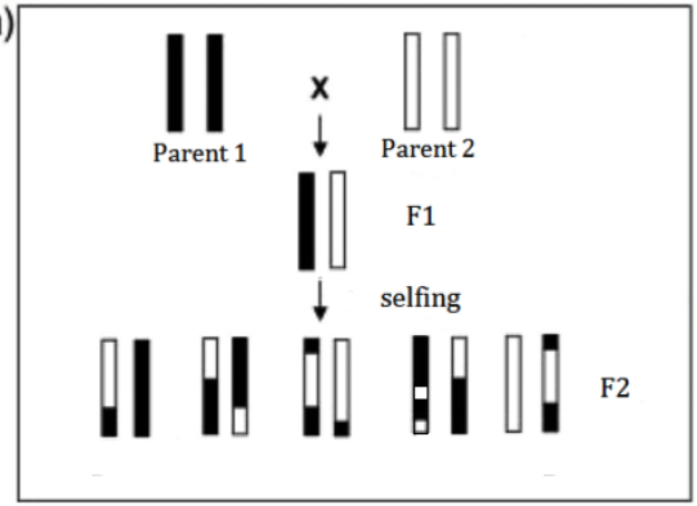
- Proposed regularized Gibbs sampler EM copula,
- Proposed regularized approximated EM copula,
- Nonparanormal skeptic, NPNtau,
- Nonparanormal normal-score, NPNscore.





Process of linkage map construction

a)



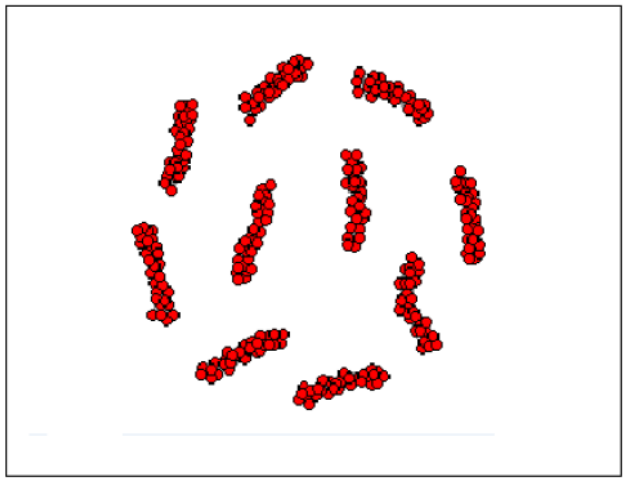
b)

	M1	M2	M3	M4	M5	M6	M7	M8	...	M500
1	0	0	0	1	1	1	2	2	...	2
2	0	0	0	0	0	-	2	2	...	2
3	2	2	2	2	1	1	0	0	...	-
4	0	-	0	1	1	1	1	-	...	0
200	2	2	2	1	1	1	1	2	...	0

AA= 0, Aa= 1, aa= 2

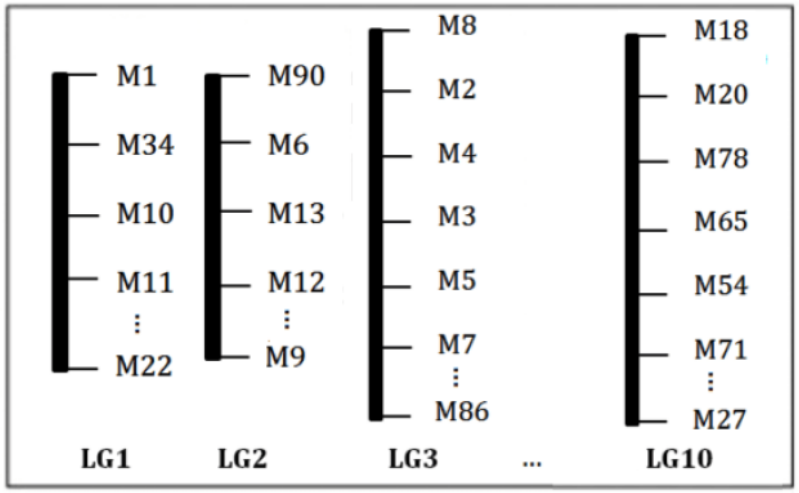
Inferring network

C)

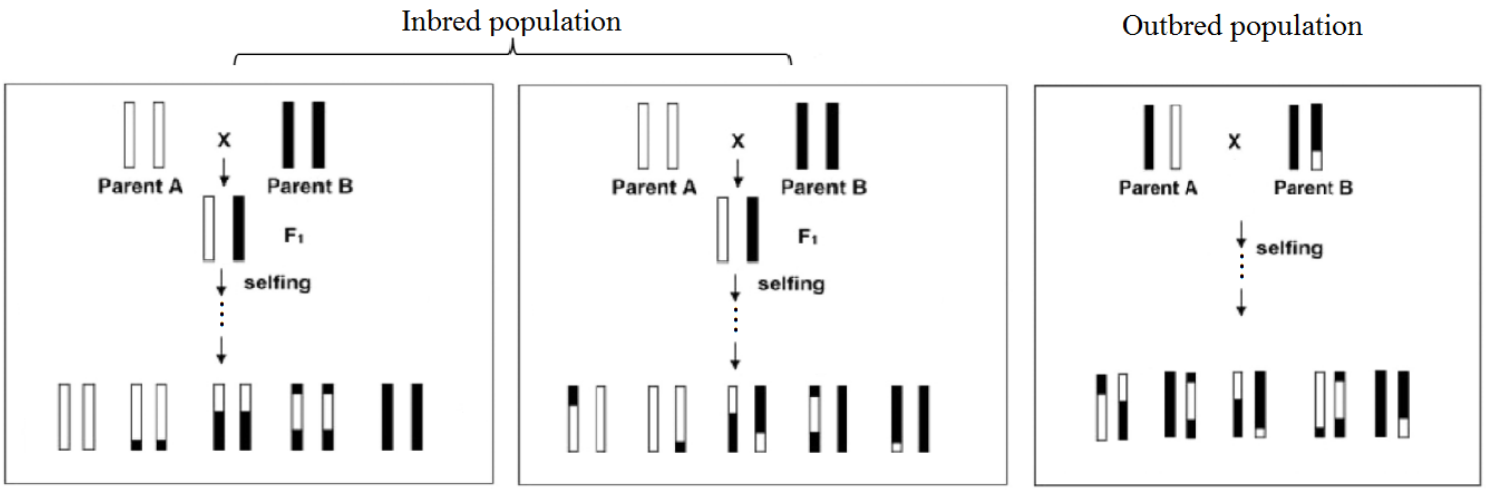


Ordering markers

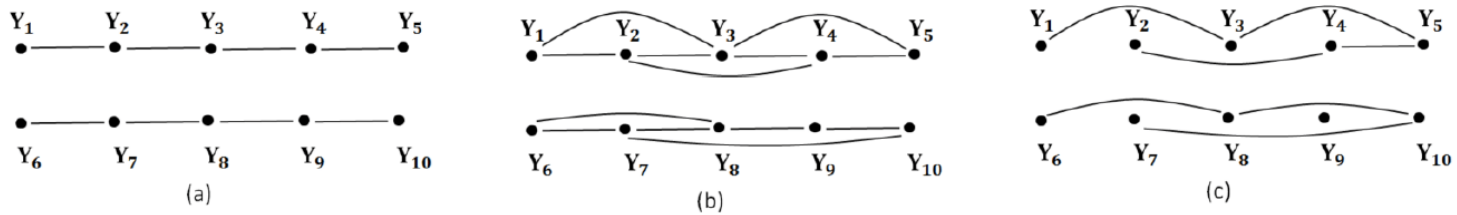
d)



Ordering adjacency matrix



Conditional dependence pattern beten neighboring markers:



Inbred populations: scheme (a) & (b)

Multi-dimensional scaling (MDS):

$$\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}}\sqrt{\theta_{jj}}}, D = -\log(\rho)$$

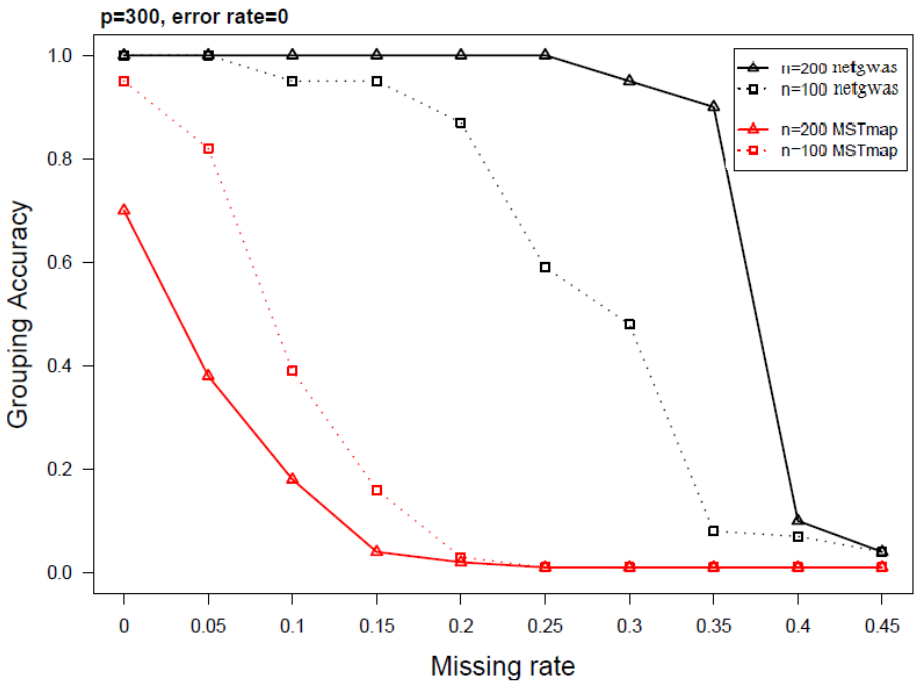
Outbred populations: scheme (c)

Reverse Cuthill-McKee (RCM) algorithm:

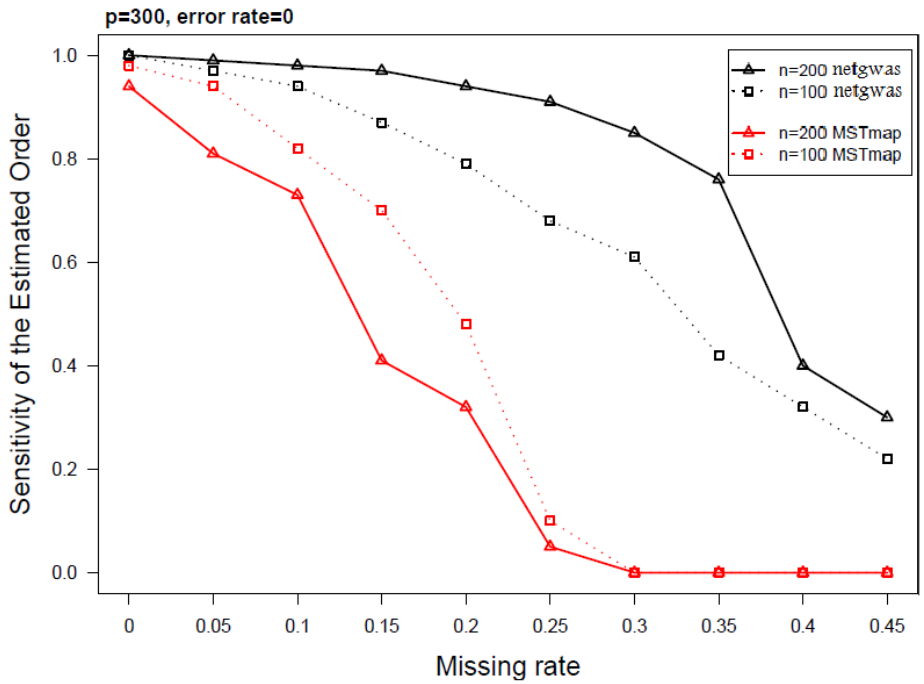
Reduces the bandwidth of adjacency matrix A by moving the non-zero elements of matrix A closer to the main diagonal.

Simulation study: **inbred** populations

- Compare `netgwas (Behrouzi and Wit(2017)) and MSTmap (Wu et.al (2008))`
- `Genotype data Simulated from PedigreeSim (a genetic software)`
- $p = 300$ `for` $n = 100$ `and` $n = 200$
- `Different ranges of missingness`



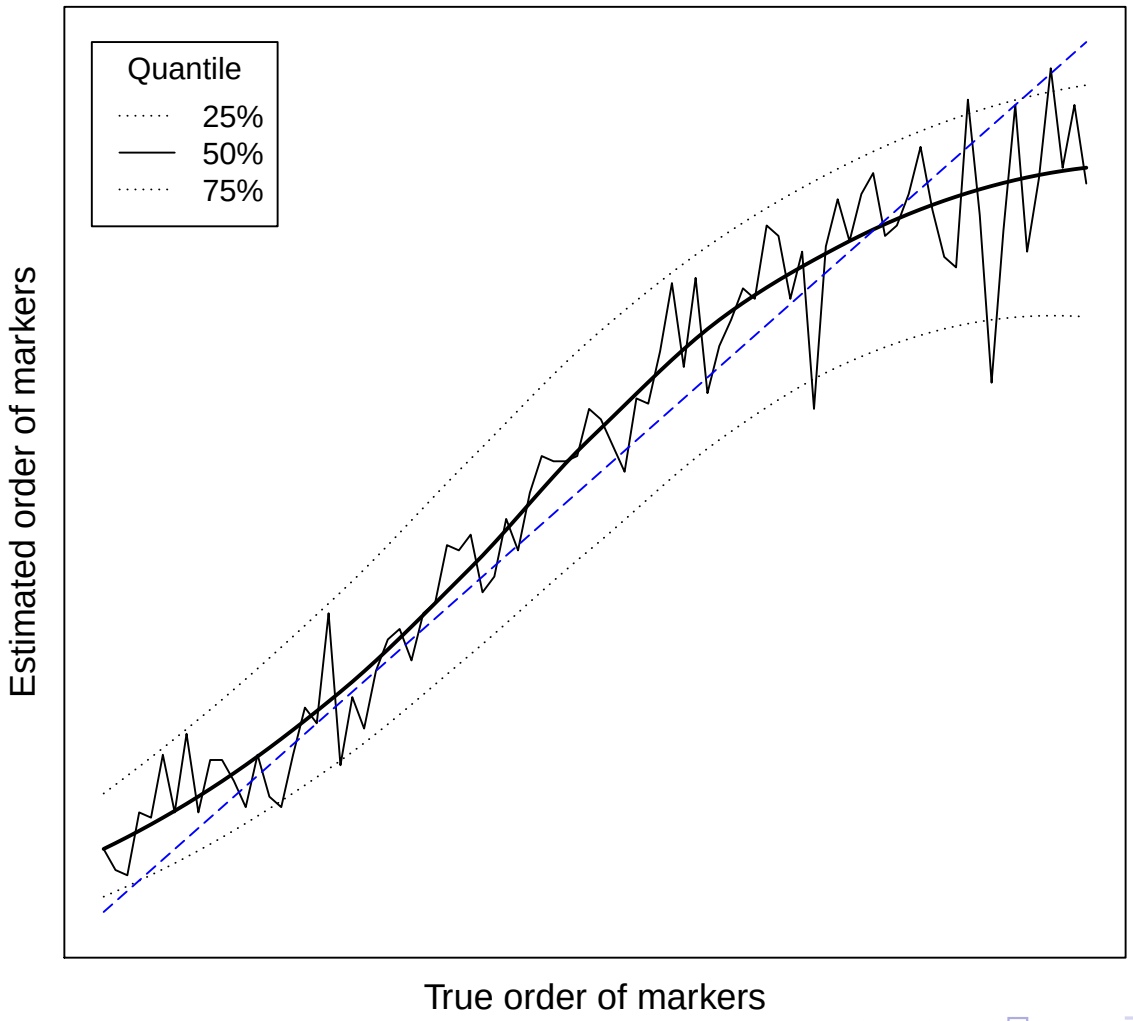
(a)
Grouping accuracy



(b)
Ordering accuracy

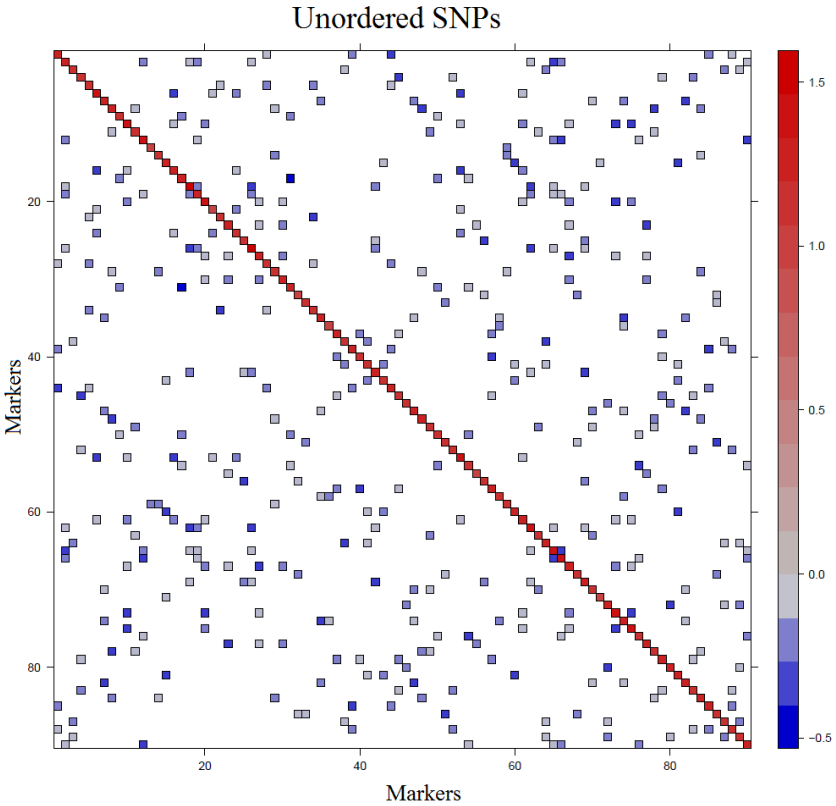
Simulation study: **outbred** populations

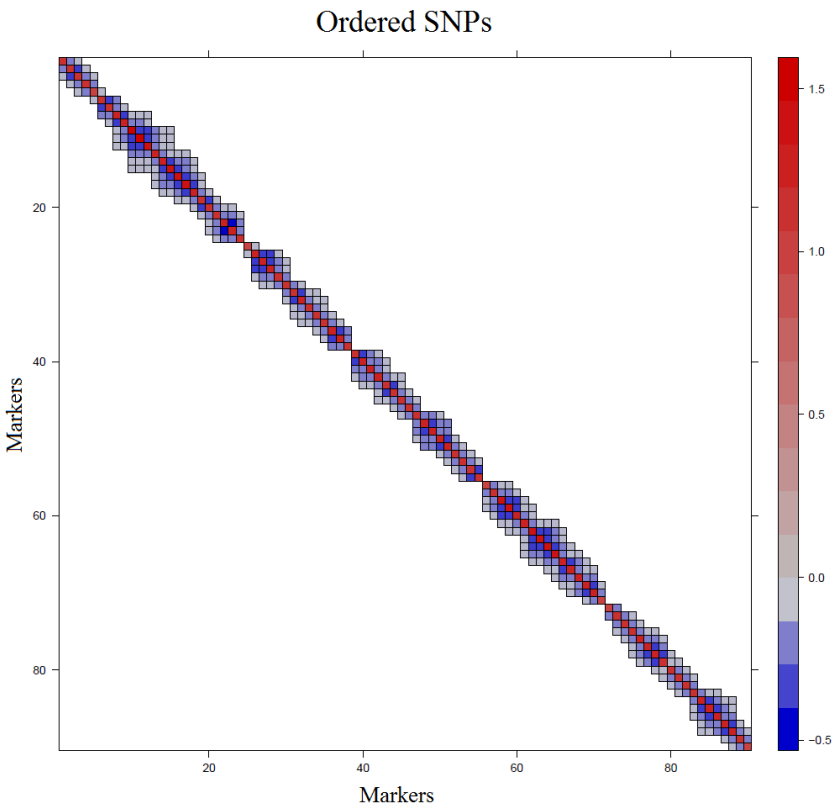
- Compare `netgwas map with the true map`
- `Genotype data simulated from PedigreeSim (a genetic software)`
- $p = 1000,\ n = 200$



Construct linkage map for **A.thaliana**

- Columbia (Col-0) and Cape Verde Island (Cvi-0)
- **Inbred** population, $Y_j^{(i)} \in \{0, 1, 2\}$
- $\boldsymbol{p} = \mathbf{90}$ SNP markers, $\boldsymbol{n} = \mathbf{367}$ individuals

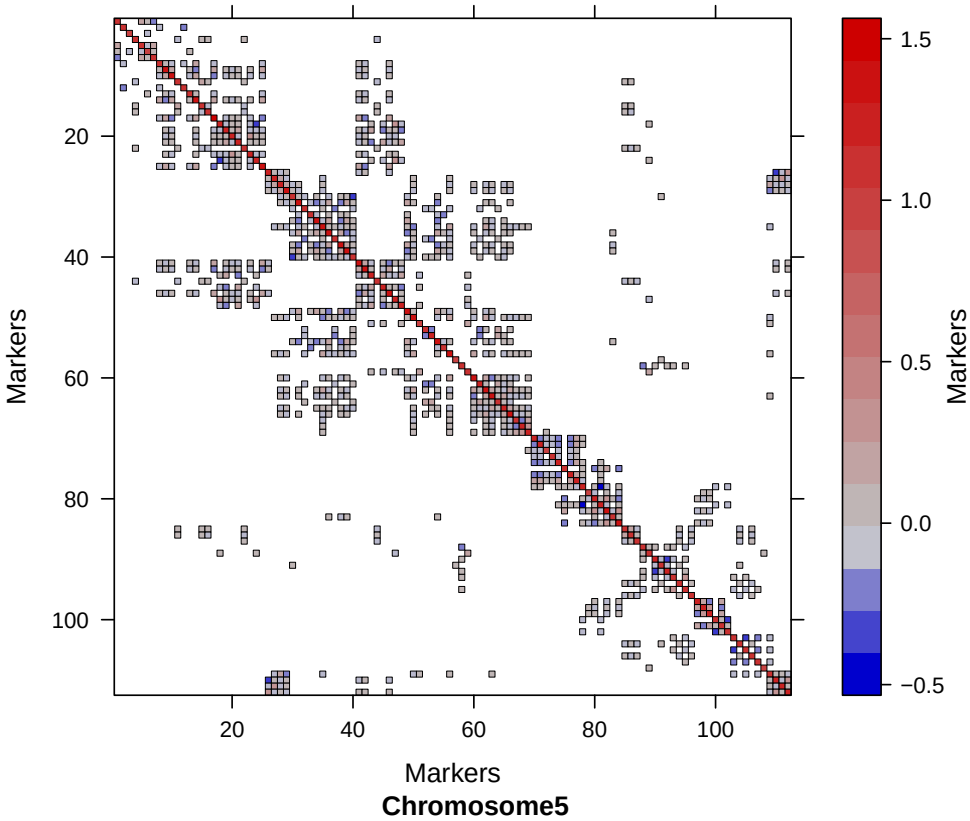




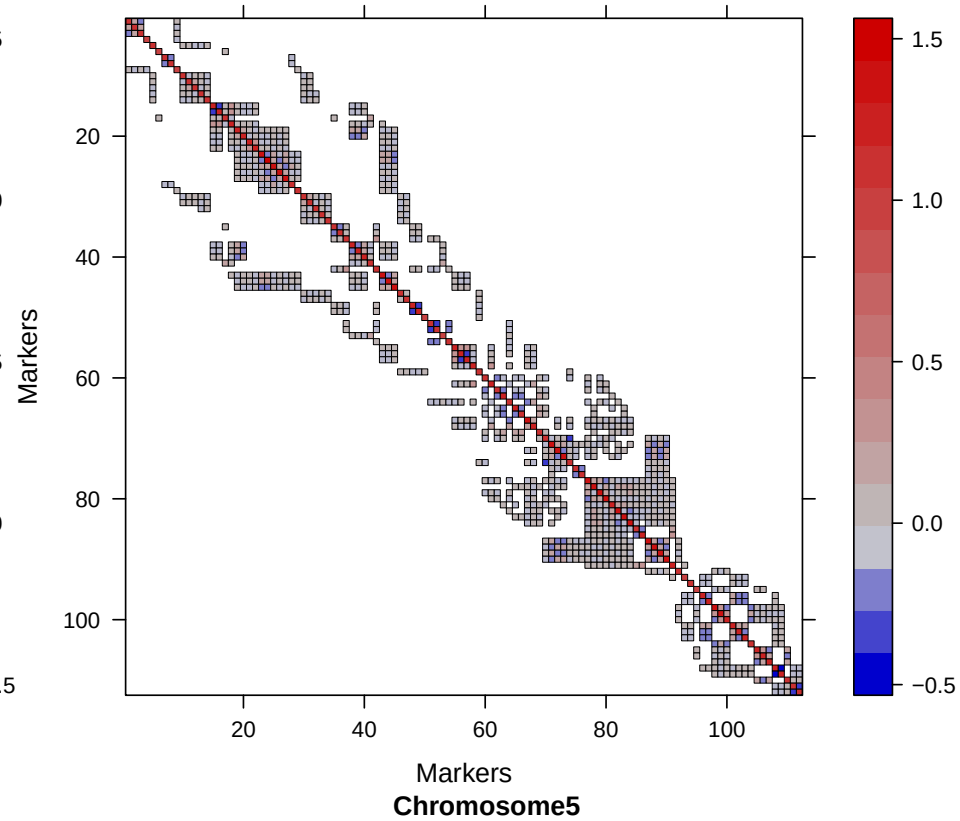
Construct linkage map for **potato**

- **Outbred** population, $Y_j^{(i)} \in \{0, 1, \ldots, 4\}$
- $\boldsymbol{p = 1972}$ SNP markers, $\boldsymbol{n = 156}$ individuals

Reverse Cuthill-McKee ordering algorithm:



$\widehat{\Theta}_\lambda$ before ordering



$\widehat{\Theta}_\lambda$ after ordering

References

Extention

Extending the method for (un)bounded discrete data, where marginals are allowed to change for different variables.

Thank you!

P. Behrouzi, E. Wit (2018). De novo construction of polyploid linkage maps using discrete graphical models. *Bioinformatics.*

P. Behrouzi, E. Wit (2019). Detecting Epistatic Selection with Partially Observed Genotype Data using Copula Graphical Models. *Royal Statistical Society (Series C)*

P. Behrouzi, D. Arrends, and E. Wit (2019). netgwas: An R package for network-based genome-wide association studies. *Submitted to JSS*

J. Guo, E. Levina, and M. George and Zhu, J (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*

T. Julian, B David (2017). R Package ASMap: efficient genetic linkage map construction and diagnosis. *JSS*