# Presentation of activities

Marie Denis

May 11, 2023




This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

May 11, 2023

# Outline

# Objectives

To present statistical models developed in response to various biological questions

1. How to detect QTLs in complex populations while considering high-dimensional data?
2. How to identify environmental variables as well as time periods involved in biological process stages ?
3. How does the genetic architecture of quantitative traits evolve over time?

Other questions if we have enough time ...

1. How to consider dependence structures between variables into statistical models?
2. Which are the relationships between variables ?

# How to detect QTLs in complex populations while considering high-dimensional data?

## Biological context

To identify QTLs related to the *oil palm* production (IBD-QTL mapping context)

## Statistical model: Linear Mixed Model (LMM)

$$y = X\beta + Zu + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

- $y$ a $n \times 1$ vector of responses (ex: height),
- $X$ a $n \times p$ matrix associated to fixed effects (ex: genotype matrix), $\beta$ the $p \times 1$ vector of fixed effects (ex: marker effects),
- $Z$ a $n \times q$ matrix associated to random effects , $u \sim \mathcal{N}_q(0, \sigma_u^2 A)$, the $q \times 1$ vector of random effects (ex: additive genetic effects),

# How to detect QTLs in complex populations while considering high-dimensional data?

## Common approaches

$$y = X\beta + Zu + \varepsilon$$

with $X\beta = X_1\beta_1 + \cdots + X_p\beta_p$: genotypes to markers with their associated effects

**To select the relevant markers $\Leftrightarrow$ Which $\beta_j's \neq 0$ for $j = 1, \ldots, p$?**

- ↬ Simple linear model with multiple testing correction: does not take into account confounded effects,
- ↬ Stepwise regression: not optimal especially in presence of collinearity,
- ↬ Penalized likelihood approaches (Lasso, Elastic-Net, Ridge regressions...): genetic variation is related to SNP

# How to detect QTLs in complex populations while considering high-dimensional data?

## Proposed approach

$$y = X\beta + Zu + \varepsilon$$

with $Zu = Z_1 u_1 + \cdots + Z_m u_m$: decomposition of global genetic effect as a sum of **m** local effects related to each **position/block** considered on genome ($u_j \sim \mathcal{N}_q(0, \sigma^2_{u_j} A_j)$, $A_j$ IBD matrix) $\Rightarrow$ Idea that such matrices **reflect** genetic information **better** than (single) SNP $\Rightarrow$ Regional heritability mapping

**To select the relevant positions/blocks $\Leftrightarrow$ Which $\sigma^2_{u_j} \neq 0$ for $j = 1, \ldots, m$?**

- Bayesian approach for selecting fixed and random effects

  `https://github.com/Heuclin/variance_component_selection`, Heuclin et al. 2022, in revision.

# How to detect QTLs in complex populations while considering high-dimensional data?

## Application

- Population of 144 palm trees (73% La Mé, 15% Yangambi, and 3% from their combination) and 1,007 IBD matrices (grid of 3cM)

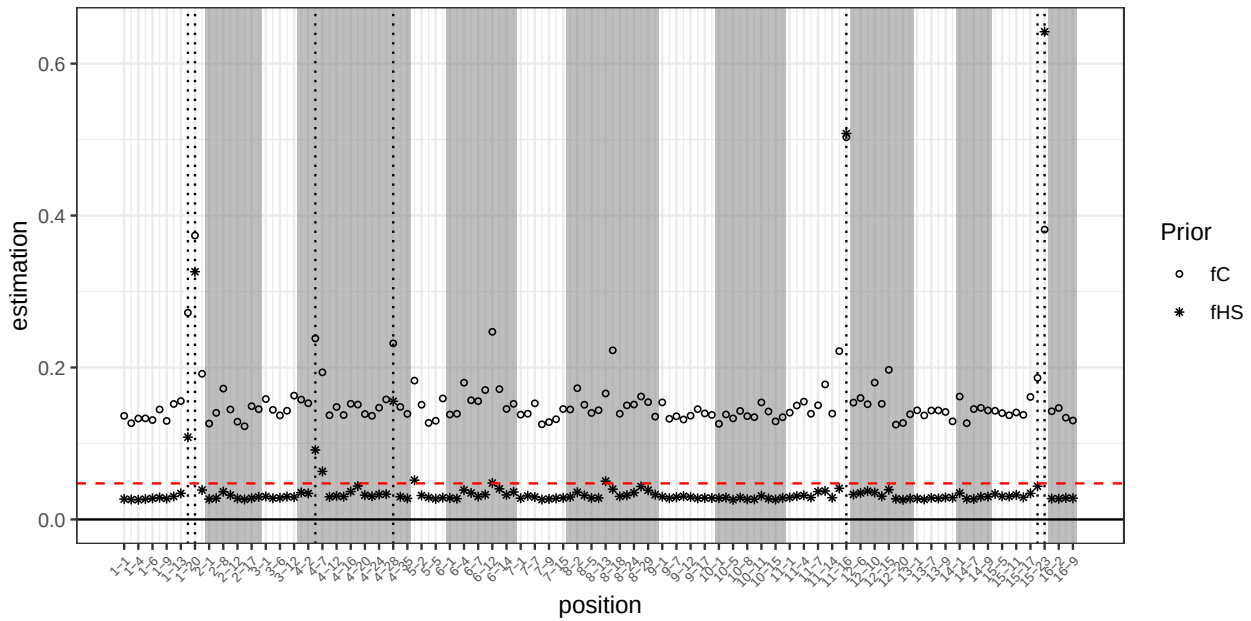



Figure 1: Posterior median of standard deviation parameters.

# How to detect QTLs in complex populations while considering high-dimensional data?

## Application

- **Results compared to other approaches**: Identification of positions corresponding to YBI QTL segregating in a minor fraction of population

↬ Performs well in unbalanced genetic designs and rare allele segregations: increases the power of detection of subtle effects

↬ Performs well even with a high number of markers

Internship of Khadidiatou Demba: to study the performances of the proposed approach as well as simple/multiple approaches in different populations (bi-parental, multi-parental)

# How to identify environmental variables as well as time periods involved in biological process stages?

## Biological context

To identify environmental variables $\boldsymbol{X_g} = (X_{g1}, \ldots, X_{gT})(g = 1, \ldots, G)$ and time periods $t \in ?$ affecting the oil palm fruit abscission process $(Y)$

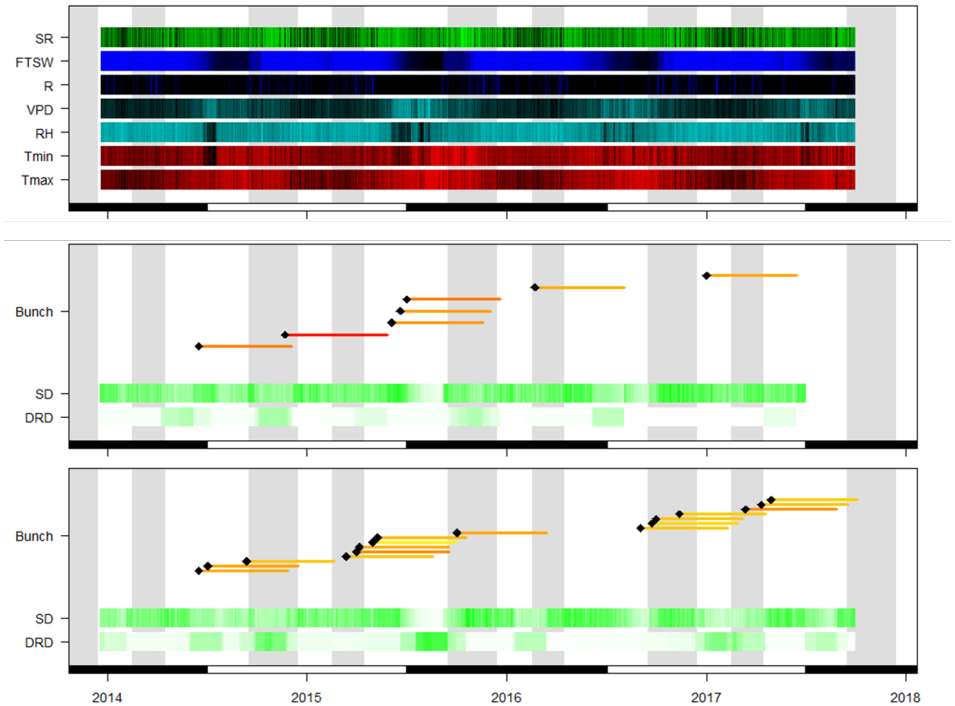



Figure 1: Representation of data over time

# How to identify environmental variables as well as time periods involved in biological process stages?

## Common approaches

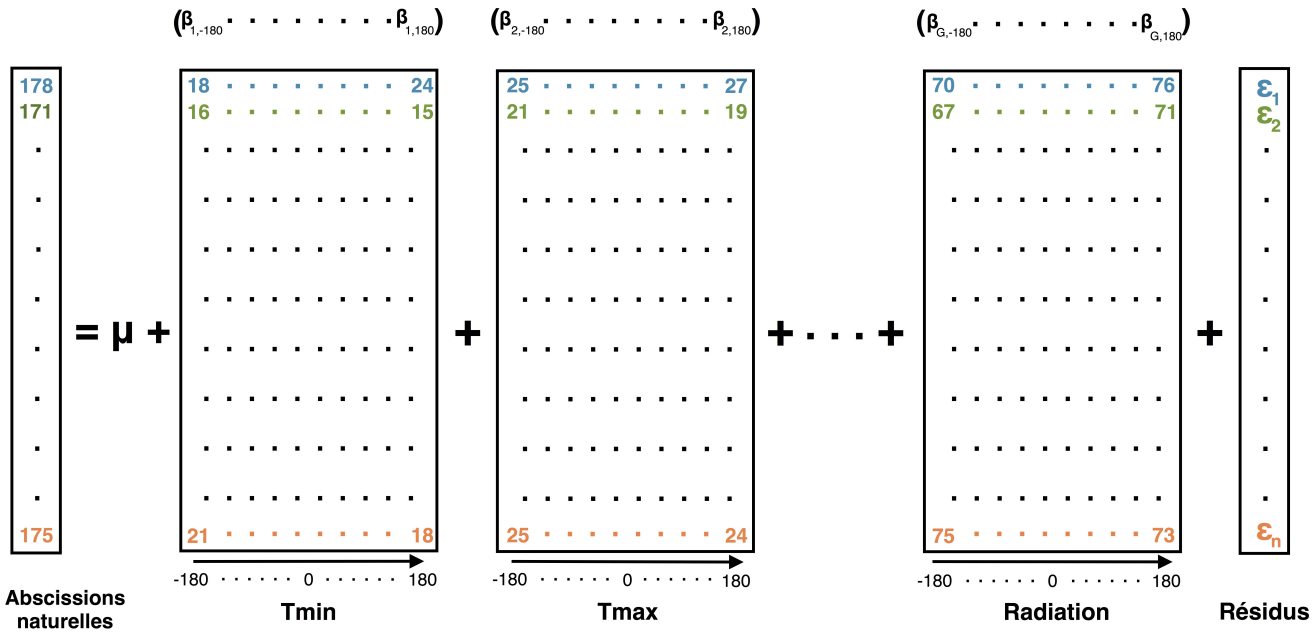
- To compute correlation between the outcome of interest $Y$ and the $g^{th}$ environmental variable $\boldsymbol{X_g}$ at each time point: $\mathrm{cor}(Y, X_{gt})$ for $t = 1, \ldots, T$,
- To perform a linear model for explaining $Y$ by $X_t$

# How to identify environmental variables as well as time periods involved in biological process stages?

## Statistical model

$$y = \mu + \sum_{g=1}^{G} \boldsymbol{X_g}\boldsymbol{\beta_g} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

- $y$ a $n \times 1$ vector of phenotype observed at one time point,
- $\boldsymbol{X_g}$ a $n \times T$ matrix of one environmental variable measured at $T$ time points, $\boldsymbol{\beta_g}$ a $T \times 1$ vector of its effects over time

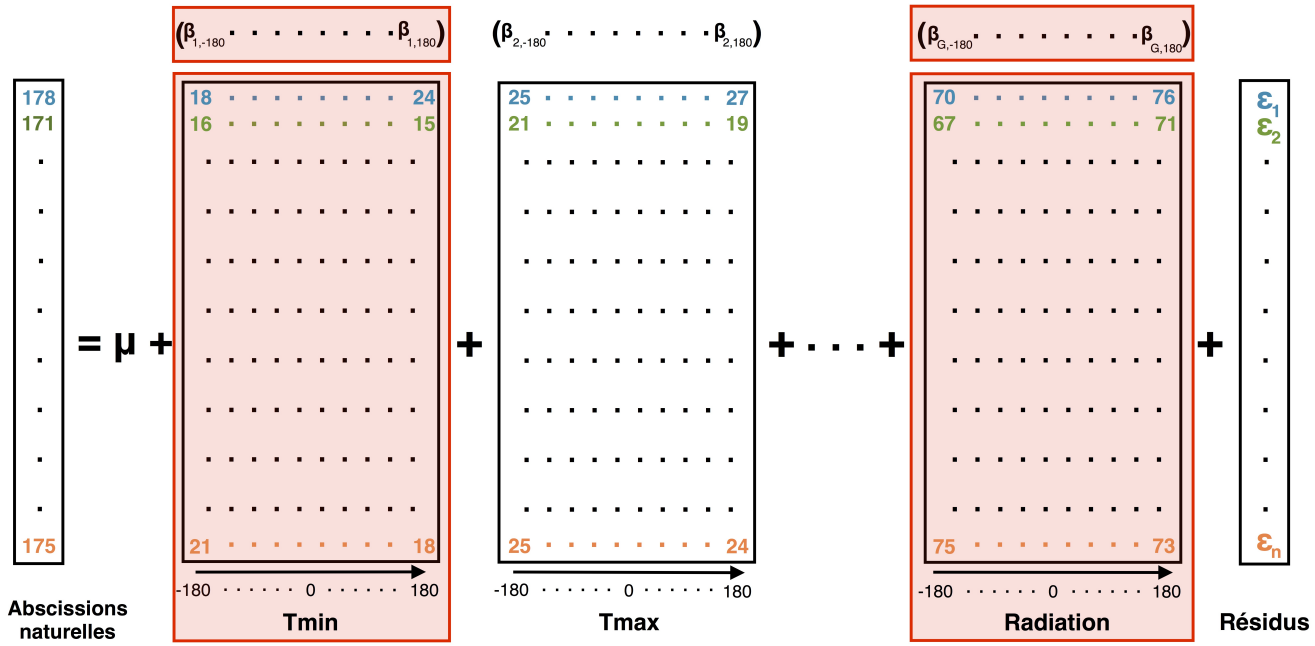


# How to identify environmental variables as well as time periods involved in biological process stages?

## "Common" approaches

- Penalized likelihood approaches considering group structure (Tisne et al. 2020)

↬ do not consider temporal structure inherent to environmental variables ($X_t$ and $X_{t+1}$ are highly correlated)

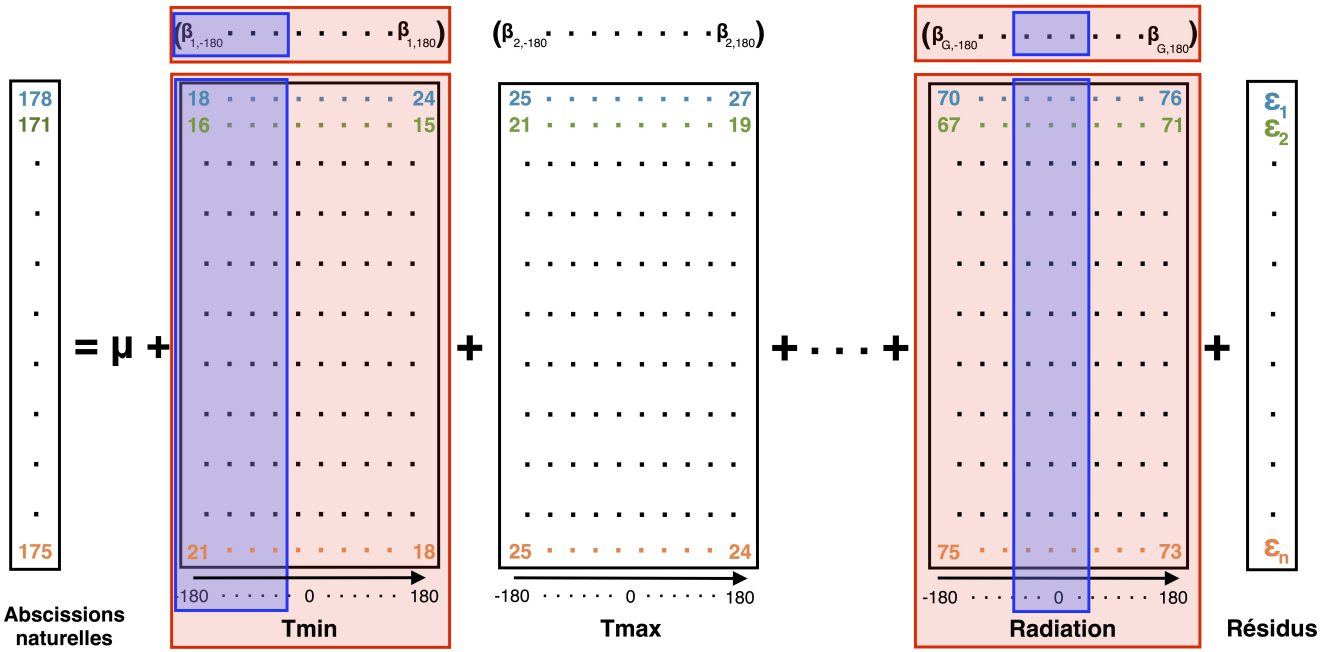


# How to identify environmental variables as well as time periods involved in biological process stages?

## Proposed approach

Bayesian variable selection approach considering dependence structure within groups of environmental variables (`https://github.com/Heuclin/GroupFusedHorseshoe`, Heuclin et al. 2022, in revision):

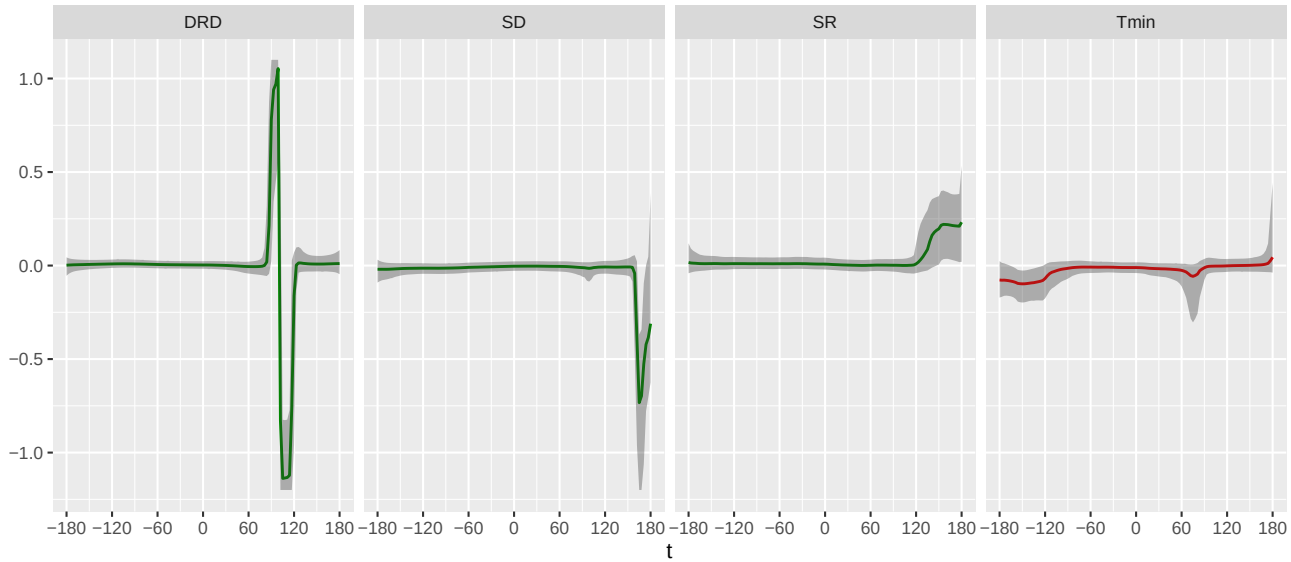
- Selection of groups of variables and time periods
- Estimation of effects over time (able to estimate simple to complex patterns over time)



# How to identify environmental variables as well as time periods involved in biological process stages?

## Application

- $1,173$ bunches, 9 environmental variables measure at $T = 121$ time points, days from pollination to fruit drop (DFD) the phenotype of interest
- **Results compared to other approaches:** Identification of 4 relevant environmental variables (with one not identified previously ) with clear profiles over time



# How does the genetic architecture of quantitative traits evolve over time?

## Biological context

To identify QTLs/markers that control the phenotypic trait over time and estimate their dynamic effects in *Arabisodpsis thaliana*

## Statistical model

$$y_t = \mu + X_j\beta_{jt} + \varepsilon_t \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2 I_n), \quad \text{for } t = 1, \ldots, T; j = 1, \ldots, J$$

- $y_t$ a $n \times 1$ vector of phenotype observed at time $t$,
- $X_j$ genotype to marker $j$, $\beta_{jt}$ the marker effect at time $t$.

# How does the genetic architecture of quantitative traits evolve over time?

## Common approach

For each time to select the relevant markers by using a simple linear model (cf first question)

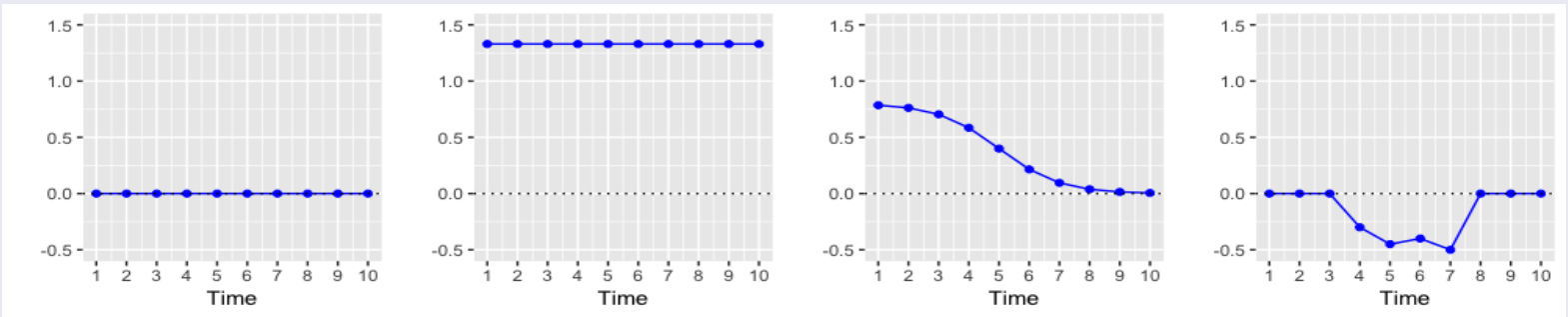
- Does not take into account dependence structure between observations measured on a same individual,
- Does not consider combined effects.

# How does the genetic architecture of quantitative traits evolve over time?

## Proposed approach

- Bayesian variable selection approach considering all markers and all time points (Heuclin et al. 2020, JRSS-C, R package VCGSS is available on github: `https://github.com/Heuclin/VCGSS`)

$$\begin{pmatrix} y_{i,t_1} \\ \vdots \\ y_{i,t_T} \end{pmatrix} = \begin{pmatrix} \mu_{t_1} \\ \vdots \\ \mu_{t_T} \end{pmatrix} + \begin{pmatrix} \beta_{1,t_1} & \cdots & \beta_{q,t_1} \\ \vdots & & \vdots \\ \beta_{1,t_T} & \cdots & \beta_{q,t_T} \end{pmatrix} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,t_1} \\ \vdots \\ \varepsilon_{i,t_T} \end{pmatrix}, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2 \Gamma) \\ \Gamma_{t,t'} = \rho^{|t-t'|} \\ -1 < \rho < 1 \end{array}$$



# How does the genetic architecture of quantitative traits evolve over time?

## Application

- 358 individuals, 125 markers, $T = 21$ time points, compactness the trait of interest
- **Results compared to other approaches:**
  - Identification of the same regions than previous study,
  - Identification of 7 new regions,
  - Estimation of dynamic effects

# Outline

# How to consider dependence structures between variables into statistical models?

Dependence structures between variables may be induced by various factors in different applications:

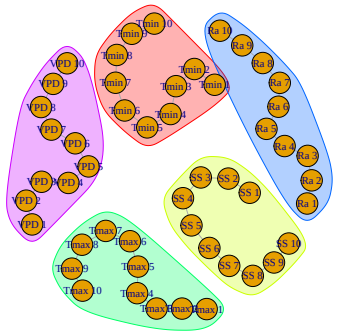
- between observations in the response variables (structure in space and/or time, structure induced by grouping factors, ...)
- between predictors (structure between genes belonging to the same biological pathway, between covariates collected over years, ..)

↬ Need to be taken into account into statistical models

# How to consider dependence structures between variables into statistical models?

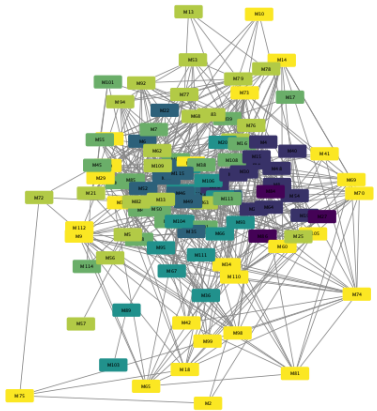
## Fruit abscission process

- nodes = climatic variables
- edges = linked when temporally dependent



## Gene network

- nodes = genes
- edges = linked when co-expressed



## Proposed approach:

Bayesian variable selection approach where the dependence structure between variables are encoded by an undirected graph (Denis and Tadesse, 2022, under revision)

# How to consider dependence structures between variables into statistical models?

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?

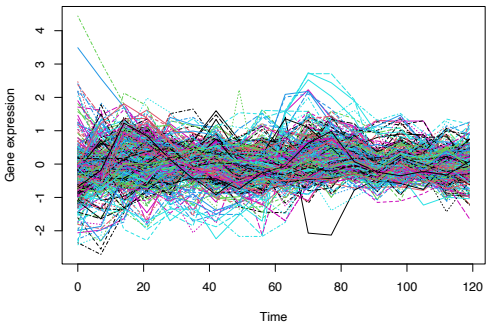



Figure 1: Cell cycle gene expression profiles over two cell cycle periods

- Which TFs are associated to the gene expression profiles? Which stage of the cell process do they influence?

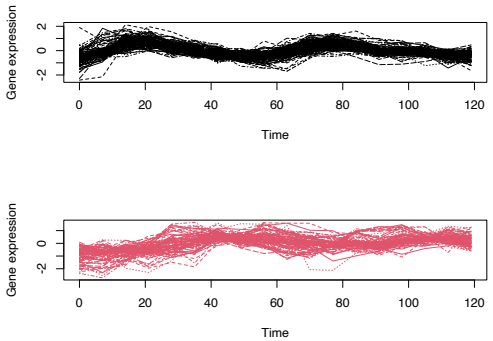



Figure 2: Gene expression profiles for two sub-groups of genes

## Proposed approach

Bayesian variable selection approach for selecting subsets of variables involved in the variability of clustered outcome profiles

# which relationships

## Biological context

Different ways for estimating gene network which are most of the time different from biological pathways:

- Pearson correlation for direct and indirect relationships,
- Partial correlation only for direct relationships,
- Adjusted partial correlation for direct relationships adjusted for other covariates (for instance miRNA - mRNA).

**Objective:** with Benedicte Favreau to estimate different networks, to understand their differences, and to compare them to biological knowledge accumulated for a process of interest. In relation with David Pot (Cresi Igen) and also in future with statisticians at INRAE and Montpellier University.

# Outline