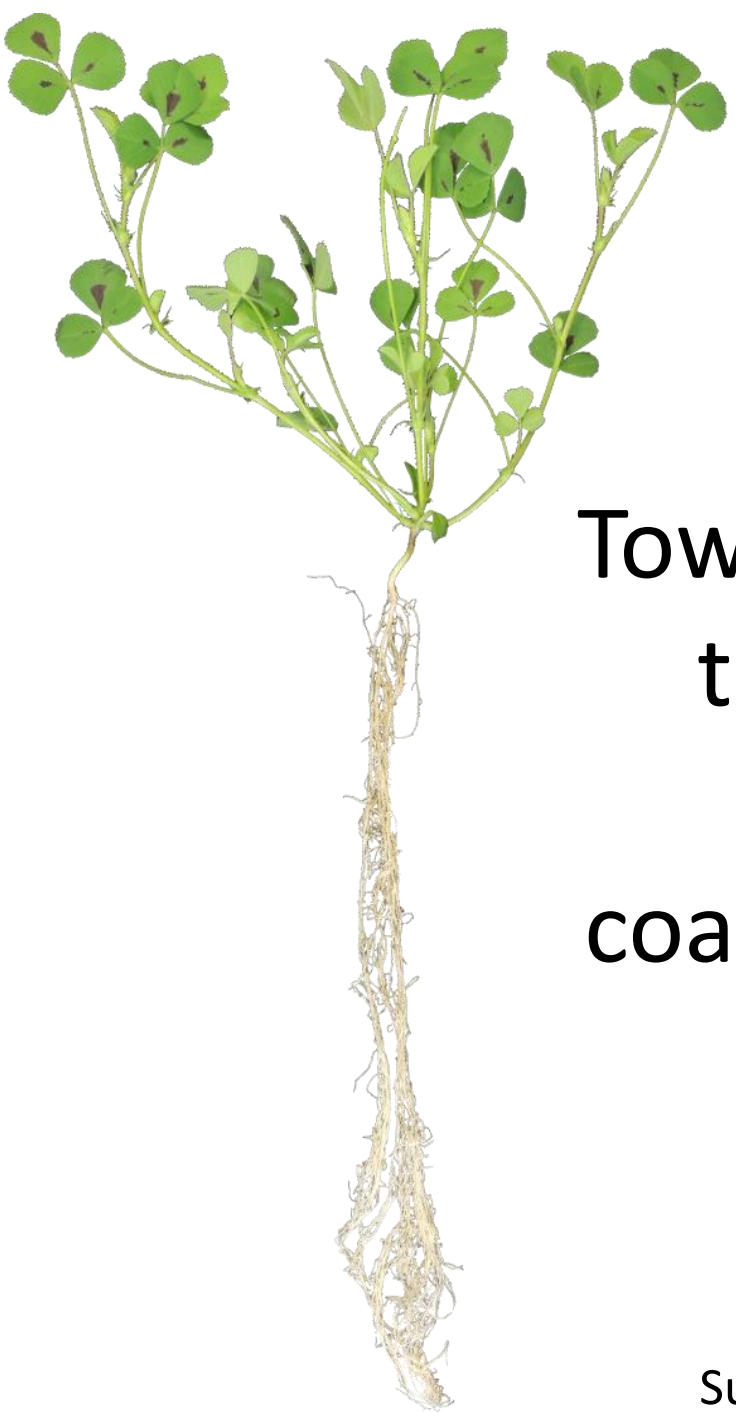


« DeCoD »
project

Towards a network approach to detect genome-wide signature of gene coadaptation using SNP data

Léa Boyrie

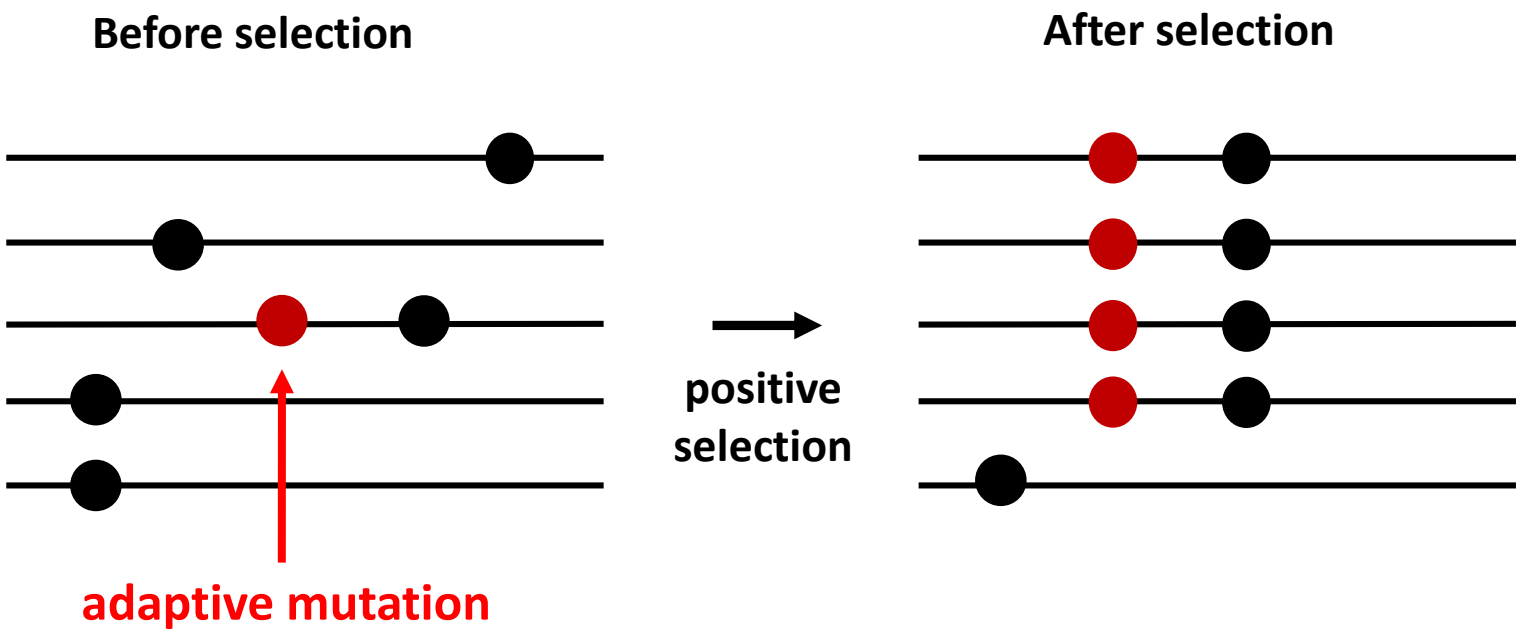
Supervisors : Maxime Bonhomme & Christophe Jacquet

Introduction

Natural selection

Induces changes in the frequency of phenotypic variants with differential fitness (survival and reproduction), and hence the genotypes/mutations associated with these phenotypic variants

Positive selection: increases the frequency of beneficial mutations over generations in the population

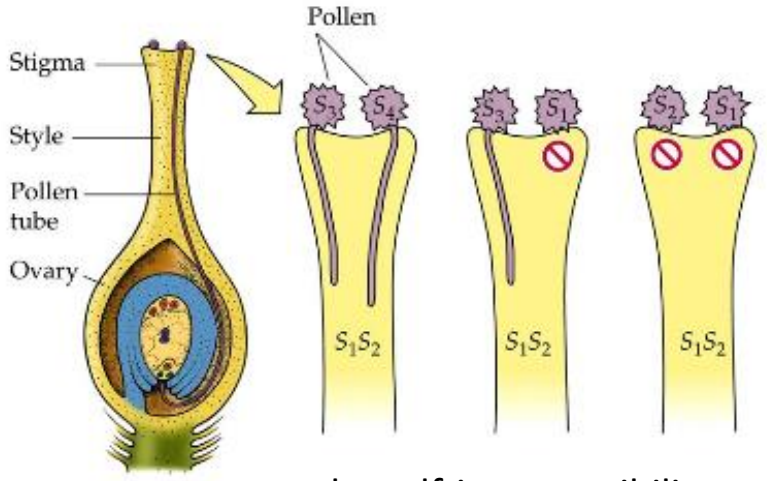


Introduction

Natural selection

Balancing selection: maintains polymorphism in a population

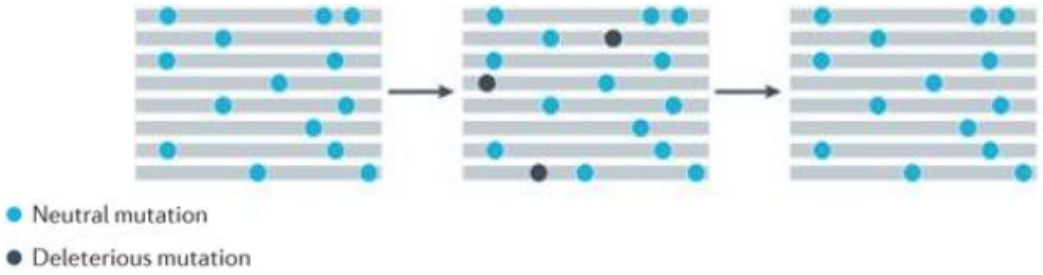
→ heterozygote advantage



example: self-incompatibility

Background Selection: deleterious alleles are eliminated by purifying selection

→ selection acting upon new deleterious mutation

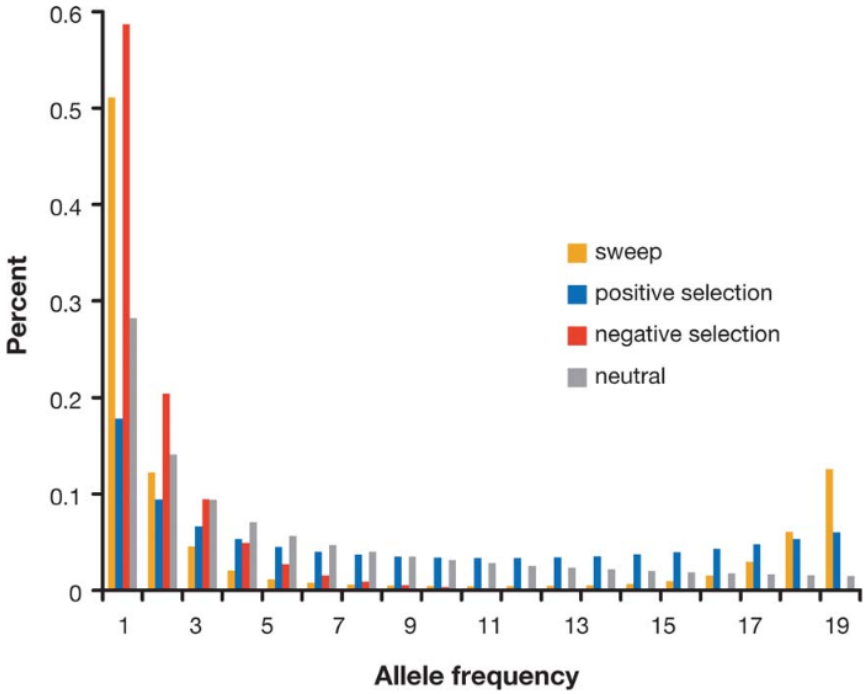


Introduction

Natural selection

Identifying the genetic bases of adaptation: methods to detect natural selection in populations

→ **Site frequency spectrum**



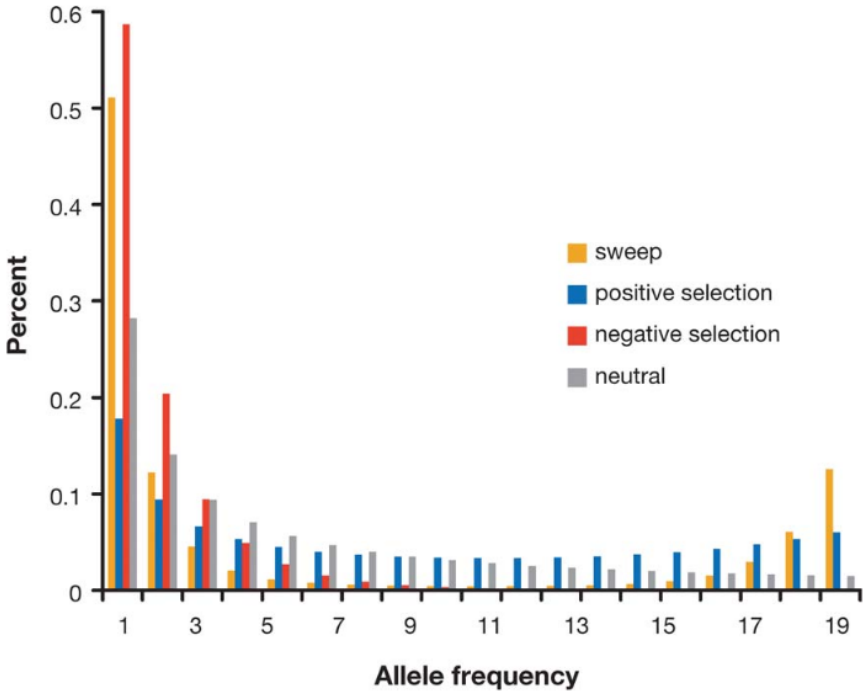
Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet.

Introduction

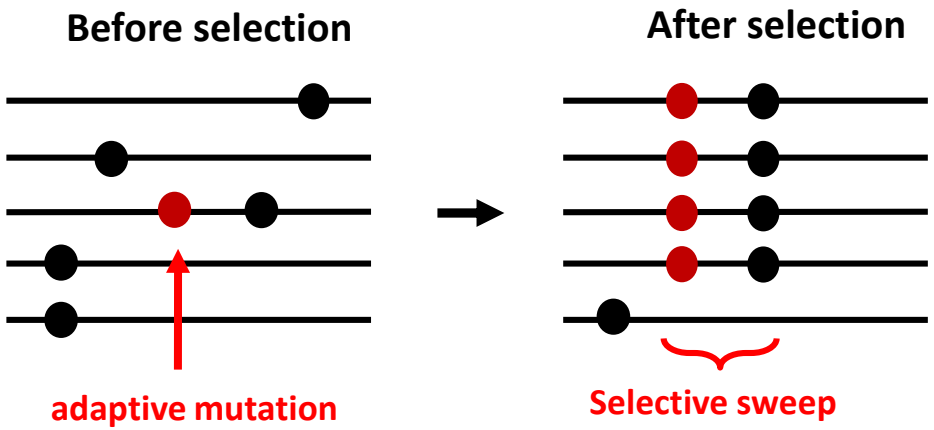
Natural selection

Identifying the genetic bases of adaptation: methods to detect natural selection in populations

→ **Site frequency spectrum**



→ **LD & Haplotype structure**



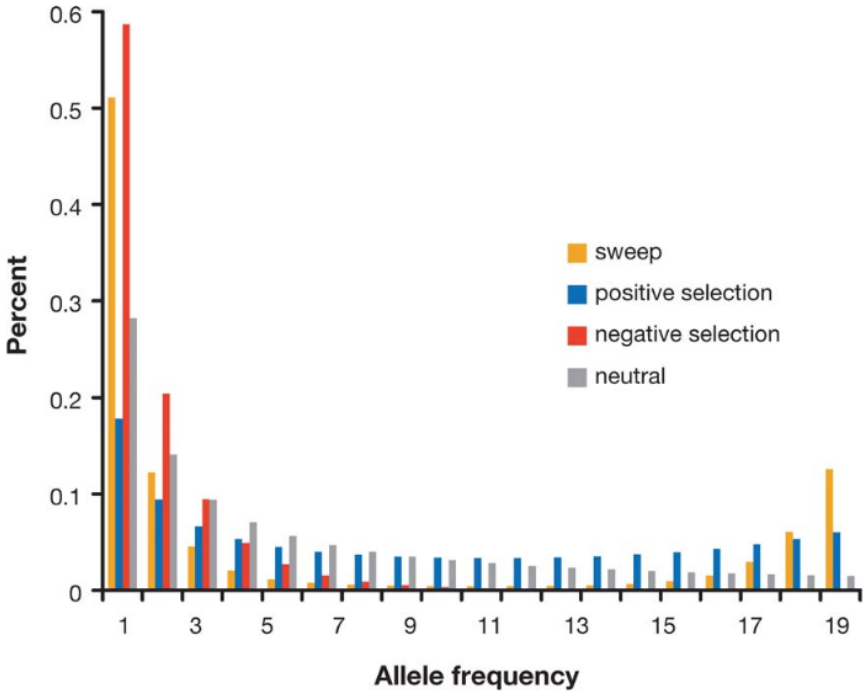
Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet.

Introduction

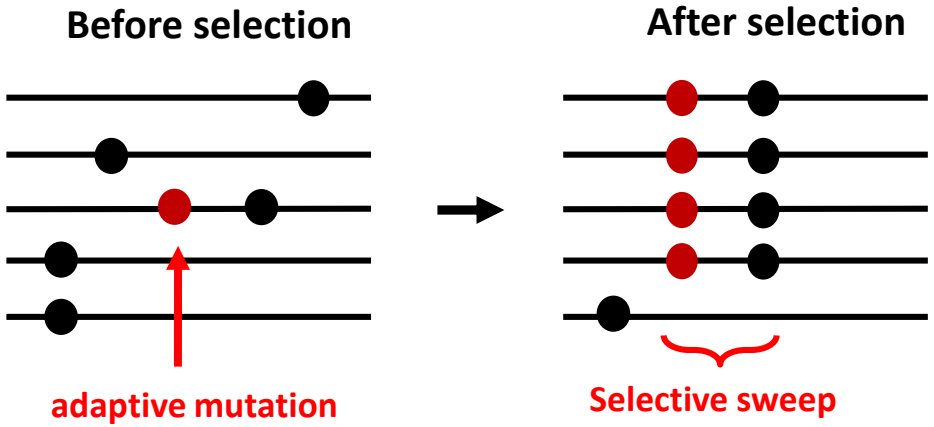
Natural selection

Identifying the genetic bases of adaptation: methods to detect natural selection in populations

→ **Site frequency spectrum**



→ **LD & Haplotype structure**



→ **Population differentiation (F_{ST})**

natural selection can causes population differentiation by local adaptation.

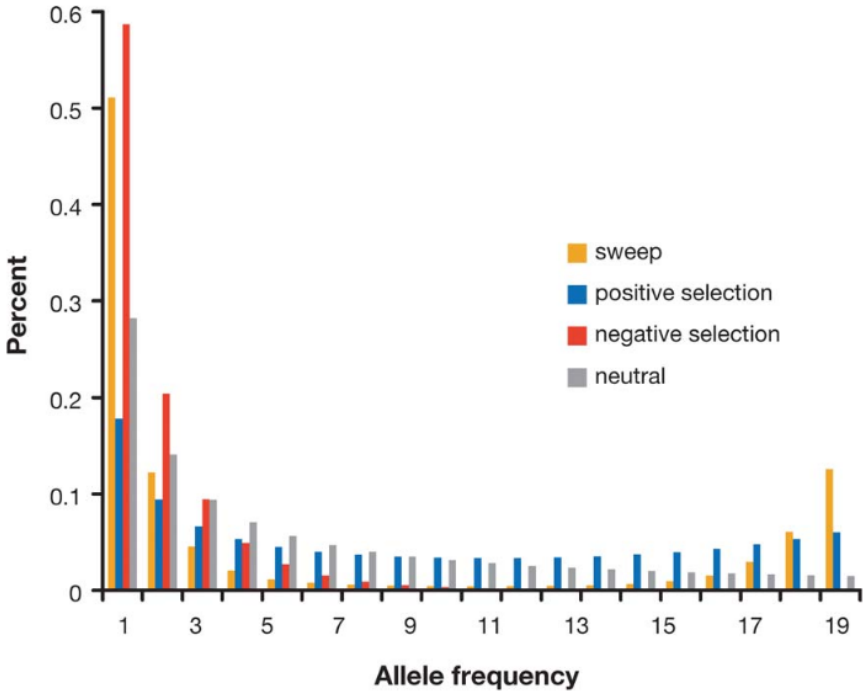
Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet.

Introduction

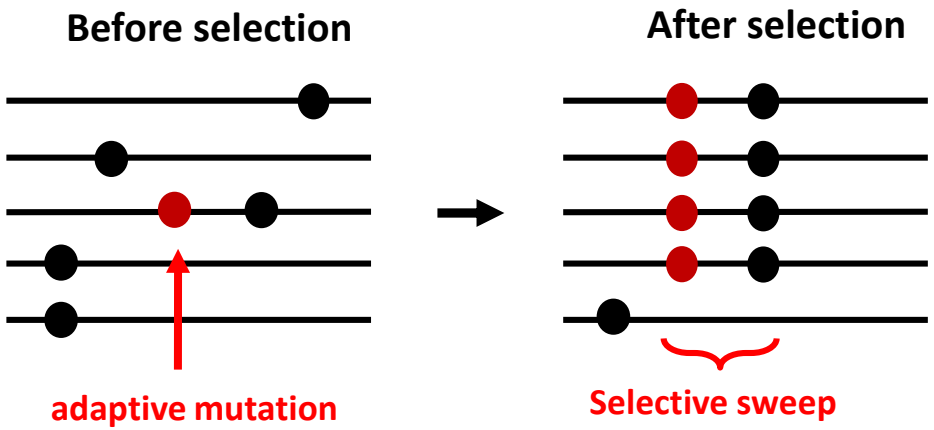
Natural selection

Identifying the genetic bases of adaptation: methods to detect natural selection in populations

→ **Site frequency spectrum**



→ **LD & Haplotype structure**



→ **Population differentiation (F_{ST})**

Natural selection can causes population differenciation by local adaptation.

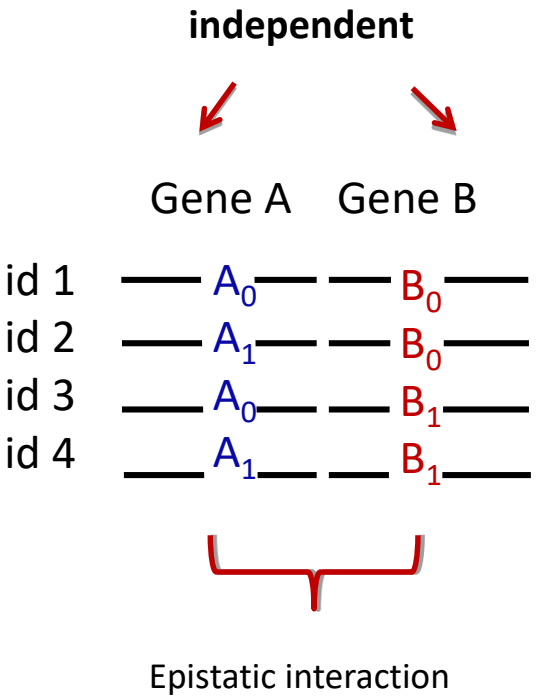
→ BUT no detection of selection acting on the interaction between genes

Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet.

Introduction

Epistatic selection

Fitness interactions among cosegregating variants (Takahasi & Innan 2008).



Haplotypes	Fitness values
A_0B_0	
A_1B_0	
A_0B_1	
A_1B_1	

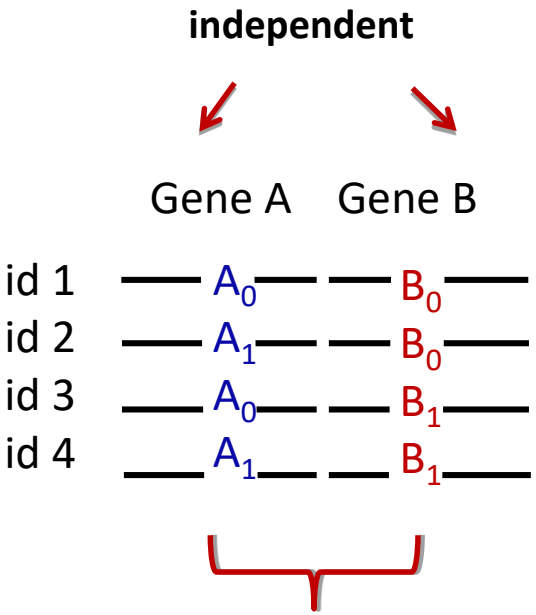
Examples: co-receptors, transcription factor complexes,…

s = 0 -> drift
s ≠ 0 -> epistatic selection

Introduction

Epistatic selection

Fitness interactions among cosegregating variants (Takahasi & Innan 2008).



Haplotypes	Fitness values
A_0B_0	1
A_1B_0	1
A_0B_1	1
A_1B_1	1+s

Coadaptation model

Two mutations A_1 and B_1 are individually neutral but together form a coadapted haplotype A_1B_1. (Takahasi & Tajima 2005)

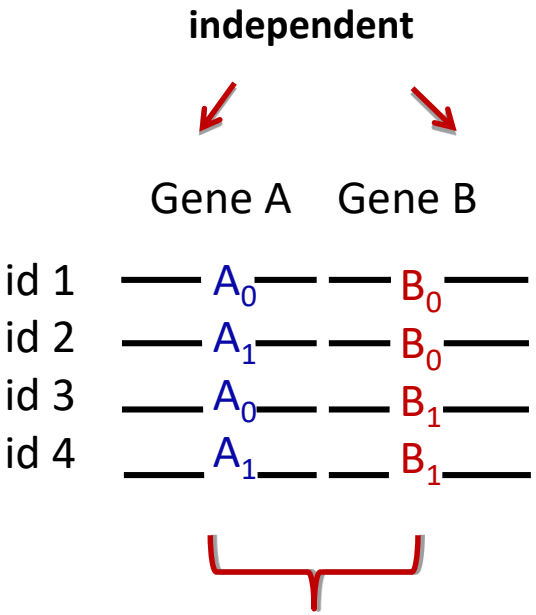
Examples: co-receptors, transcription factor complexes,…

s = 0 -> drift
s ≠ 0 -> epistatic selection

Introduction

Epistatic selection

Fitness interactions among cosegregating variants (Takahasi & Innan 2008).



Haplotypes	Fitness values	
A_0B_0	1	1
A_1B_0	1	1-s
A_0B_1	1	1-s
A_1B_1	1+s	1

Compensatory model

Two individually deleterious mutations compensate each other when combined together. (Takahasi & Innan 2008)

Examples: co-receptors, transcription factor complexes,…

s = 0 -> drift
s ≠ 0 -> epistatic selection

Introduction

Epistatic selection is detectable with linkage disequilibrium

~~LD~~

Gene A Gene B

A_0 B_0

A_1 B_0

A_0 B_1

A_1 B_1

LD

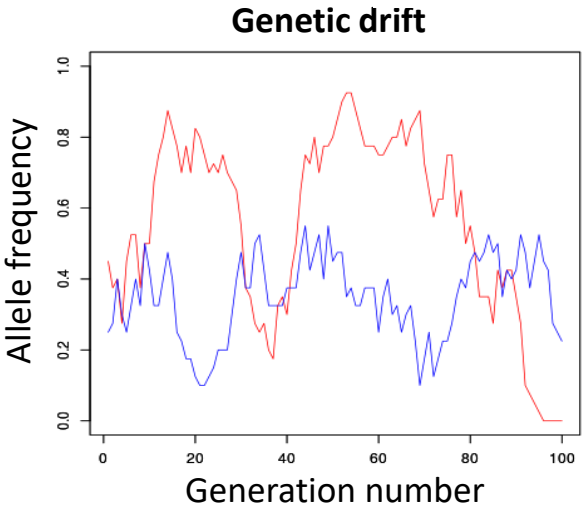
Gene A Gene B

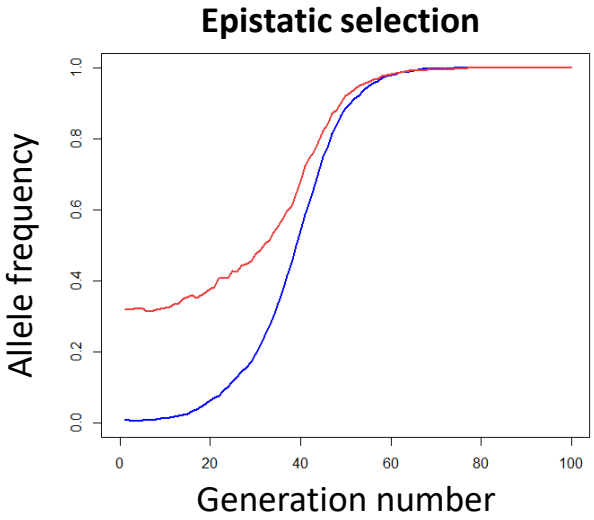
A_0 B_0

A_1 B_1

A_1 B_1

A_1 B_1





Introduction

Epistatic selection is detectable with linkage disequilibrium

~~LD~~

Gene A Gene B

A_0 B_0

A_1 B_0

A_0 B_1

A_1 B_1

LD

Gene A Gene B

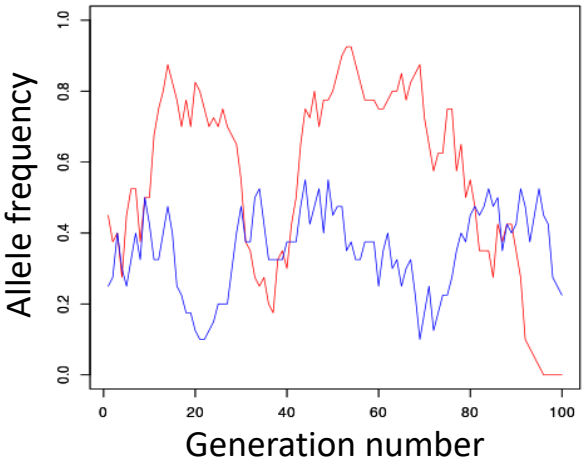
A_0 B_0

A_1 B_1

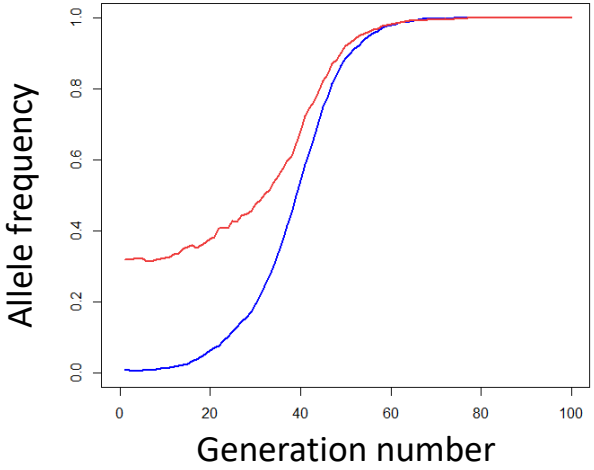
A_1 B_1

A_1 B_1

Genetic drift



Epistatic selection



measure of non random association of alleles at the two loci

Thesis Objective: **Develop a statistical test to identify genes or genomic regions in coevolution by epistatic selection signatures and find new candidates genes in association with known genes from SNP genetic data in the model legume *M. truncatula*.**

Thesis Objective: **Develop a statistical test to identify genes or genomic regions in coevolution by epistatic selection signatures and find new candidates genes in association with known genes from SNP genetic data in the model legume *M. truncatula*.**

- **Part 1: Genetic simulation and statistical detection of epistatic selection.**
  - simulation of epistatic selection
  - statistical detection of epistatic selection

- **Part 2: Detection of genes under epistatic selection in *Medicago truncatula* and in humans.**
  - SNP analyses in *Medicago truncatula:* "bait" methods
  - SNP analyses in humans: "bait" method

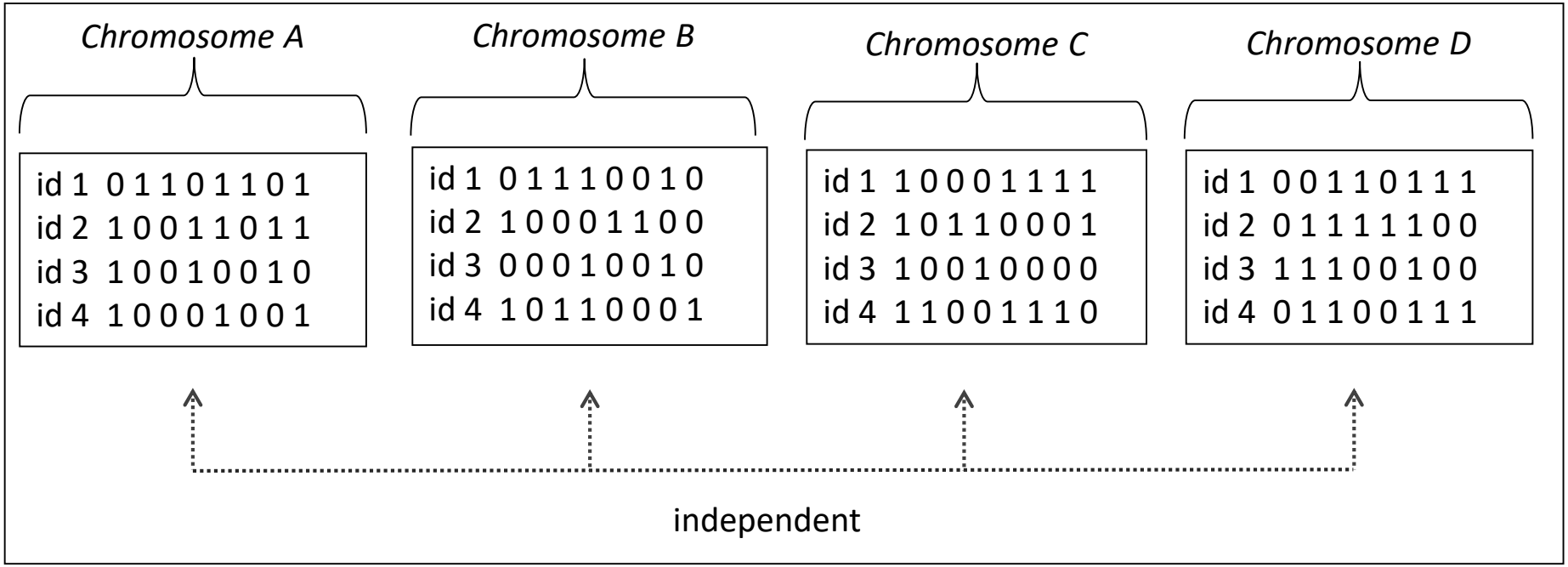
- **Part 3: Detection of coadapted clusters by genes correlation network analysis**
  - Genome-wide methods: Gene network analysis with adaptive interaction and identification of new candidate.

Simulations «*backward*»

Ancestral population

- N = 500 (diploid)
- chromosome = 5 Mb (~ 15000 marqueurs)
- 1 SNP / ~ 333 pb

- Simulation of four chromosomes

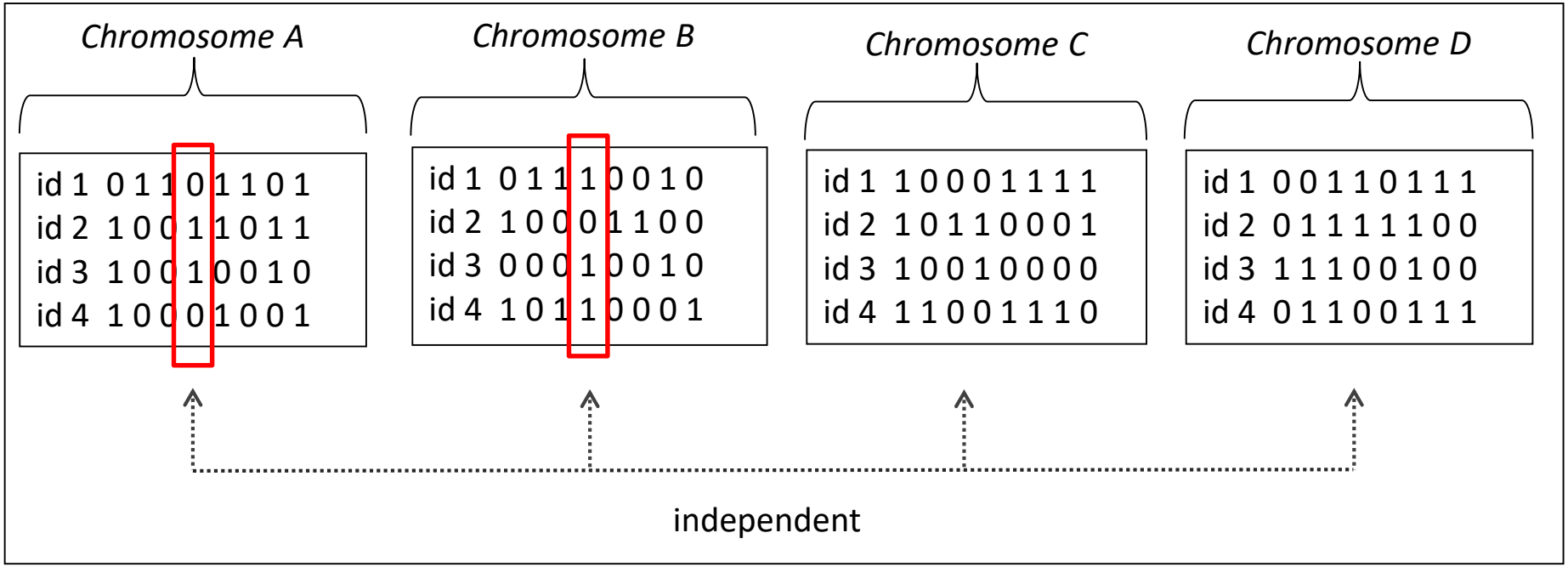


Simulations «*backward*»

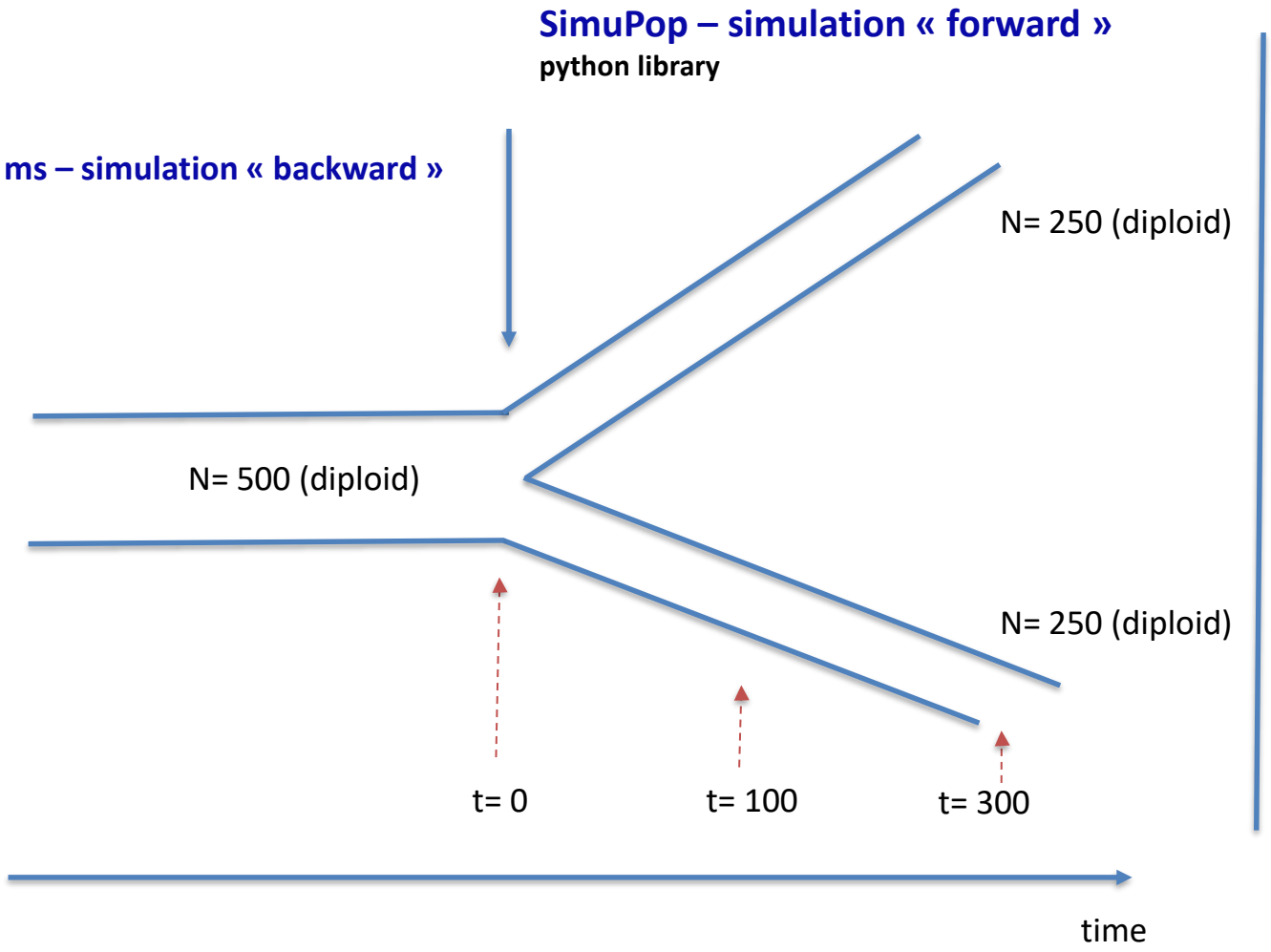
Ancestral population

- N = 500 (diploid)
- chromosome = 5 Mb (~ 15000 marqueurs)
- 1 SNP / ~ 333 pb

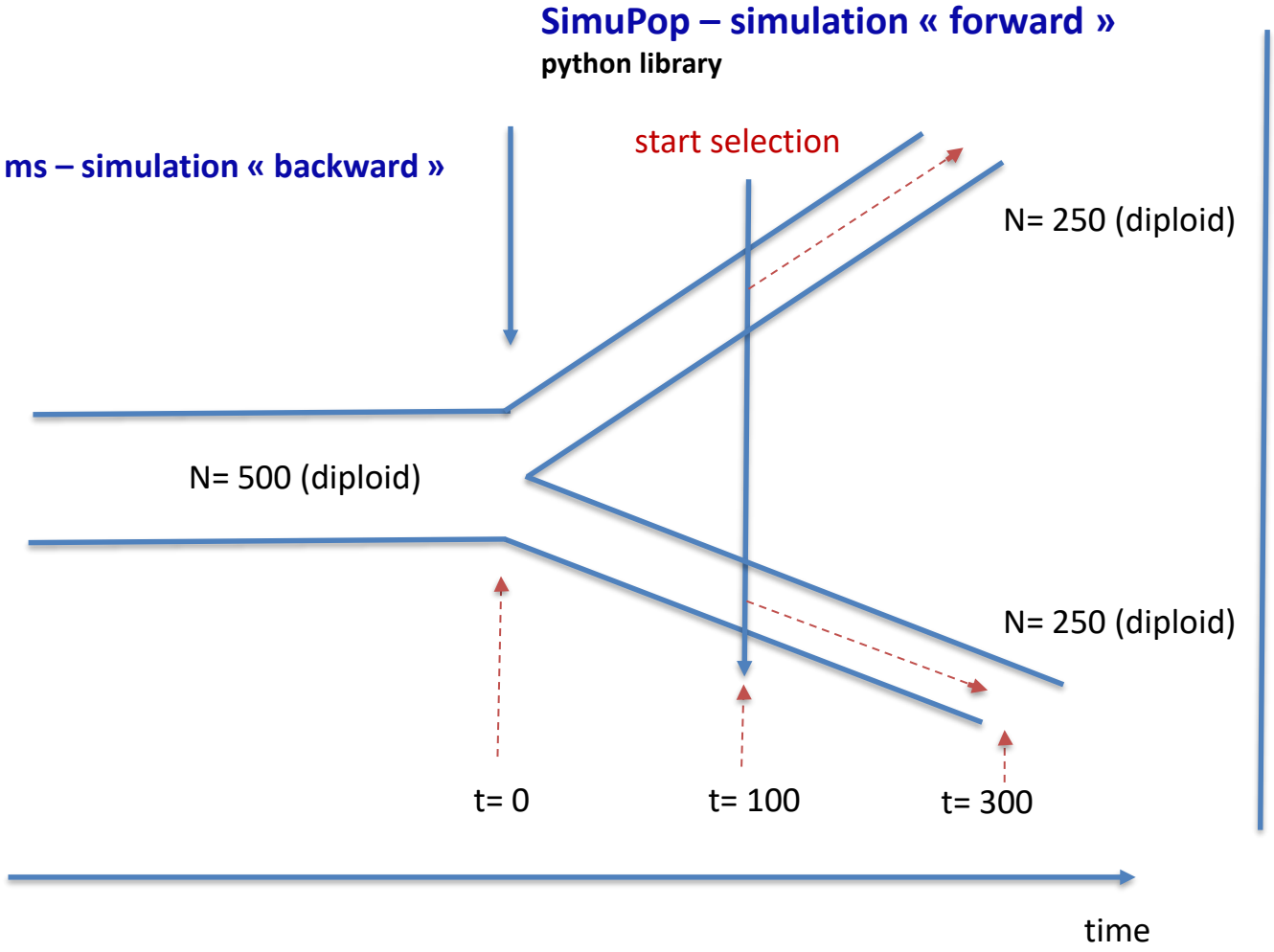
- Simulation of four chromosomes



Summary of Simulations steps



Summary of Simulations steps

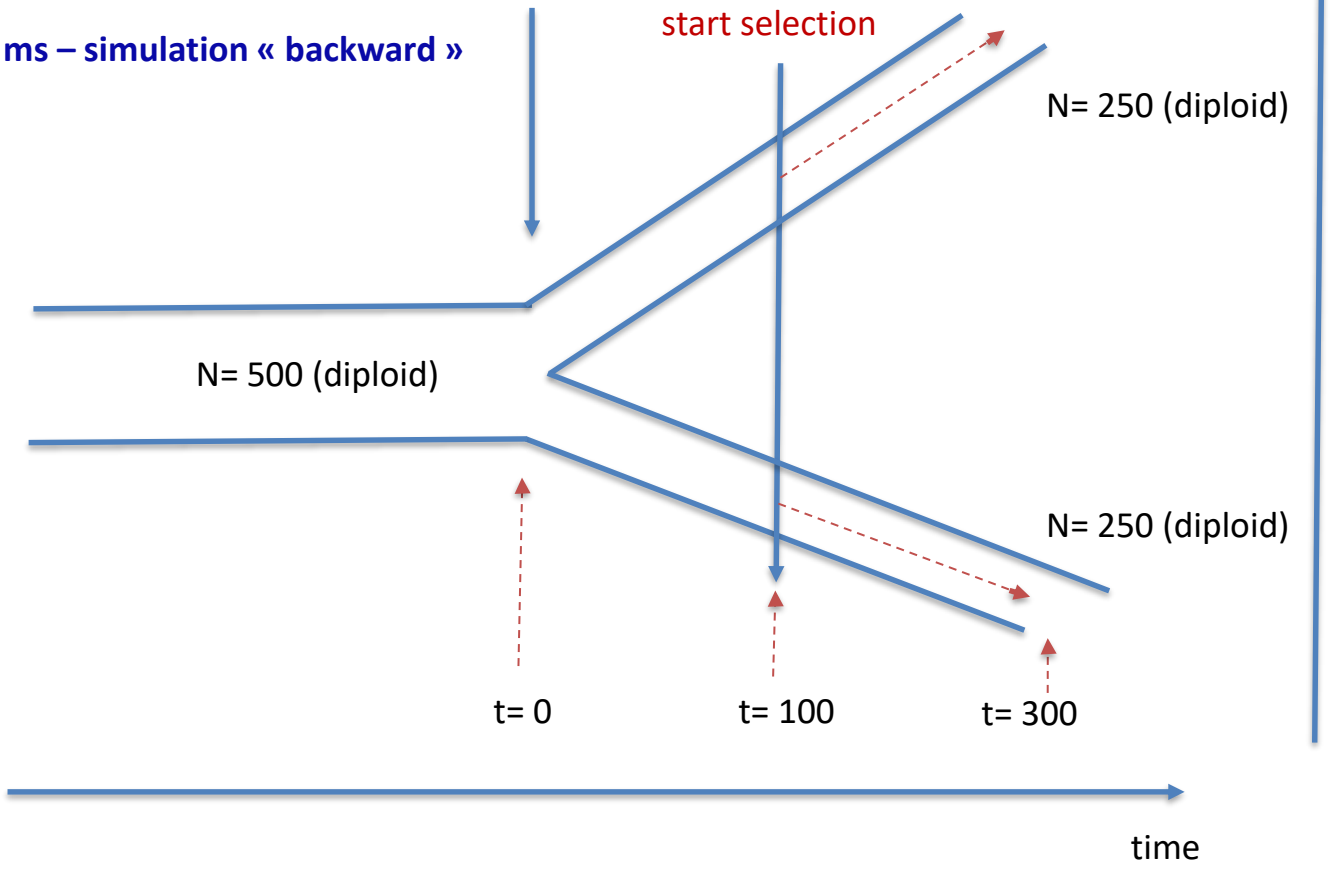


1. **Two mating schemes**
random mating, self mating

Summary of Simulations steps

SimuPop – simulation « forward »
python library

ms – simulation « backward »



2. Three selective model

Neutral	
A_0B_0	$w = 1$
A_1B_0	$w = 1$
A_0B_1	$w = 1$
A_1B_1	$w = 1$

Epistatic coadapted	
A_0B_0	$w = 1$
A_1B_0	$w = 1$
A_0B_1	$w = 1$
A_1B_1	$w = 1 + s$

Epistatic compensatory	
A_0B_0	$w = 1$
A_1B_0	$w = 1\text{-}s$
A_0B_1	$w = 1\text{-}s$
A_1B_1	$w = 1$

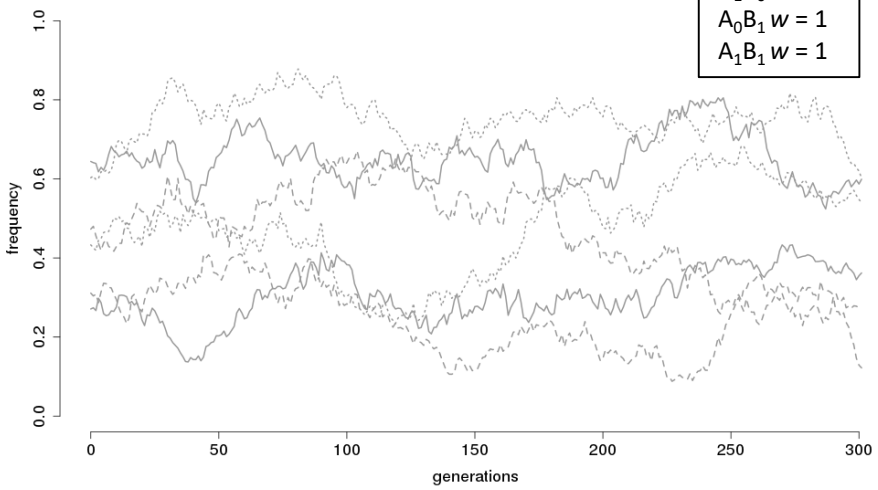
Thesis: theoretical part

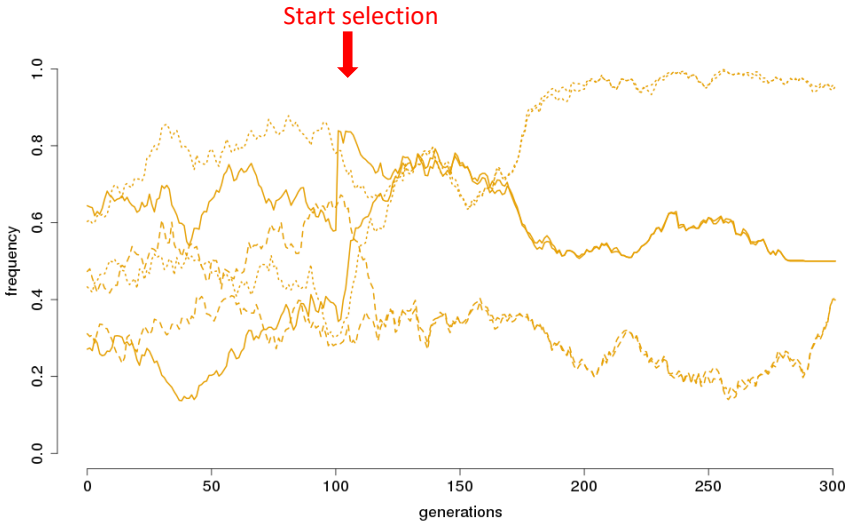
simulation 1
simulation 2
simulation 3

Neutral model
Compensatory model
Coadapted model

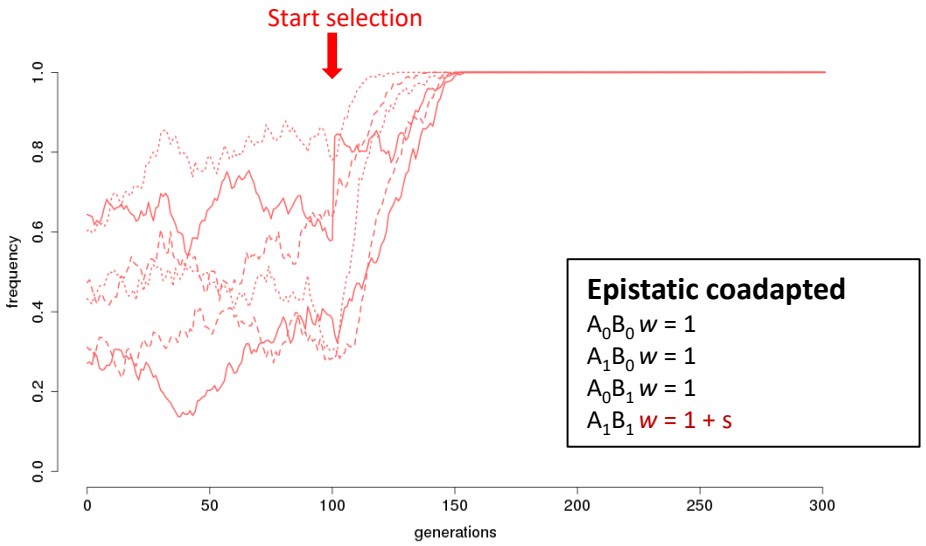
Evolution of allelic frequencies A_1 and B_1

Neutral
A_0B_0 $w = 1$
A_1B_0 $w = 1$
A_0B_1 $w = 1$
A_1B_1 $w = 1$





Epistatic compensatory
A_0B_0 $w = 1$
A_1B_0 $w = 1\text{-}s$
A_0B_1 $w = 1\text{-}s$
A_1B_1 $w = 1$

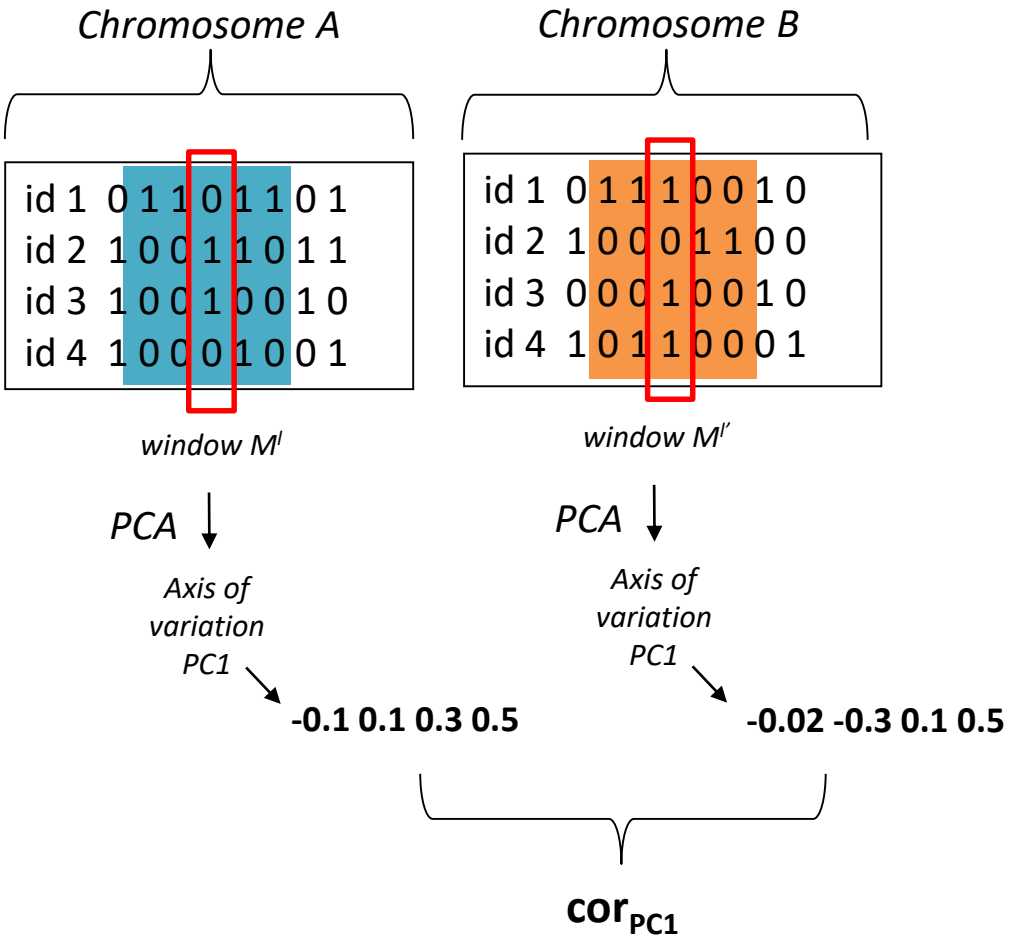


Epistatic coadapted
A_0B_0 $w = 1$
A_1B_0 $w = 1$
A_0B_1 $w = 1$
A_1B_1 $w = 1 + s$

- Neutral model: random genetic drift.
- Coadapted model: rapid fixation of the double mutant A_1B_1.
- Compensatory model: markers segregrates and alleles A_1 and B_1 (respectively A_0 and B_0) co-evolve.

Linkage disequilibrium Statistics

Haplotype calculated with PCA



Correlation between two windows

$$cor_{PC1} = cor(M^{l}_{PC1}, M^{l'}_{PC1})$$

Correction by the relatedness matrix

Correlation is biased when:

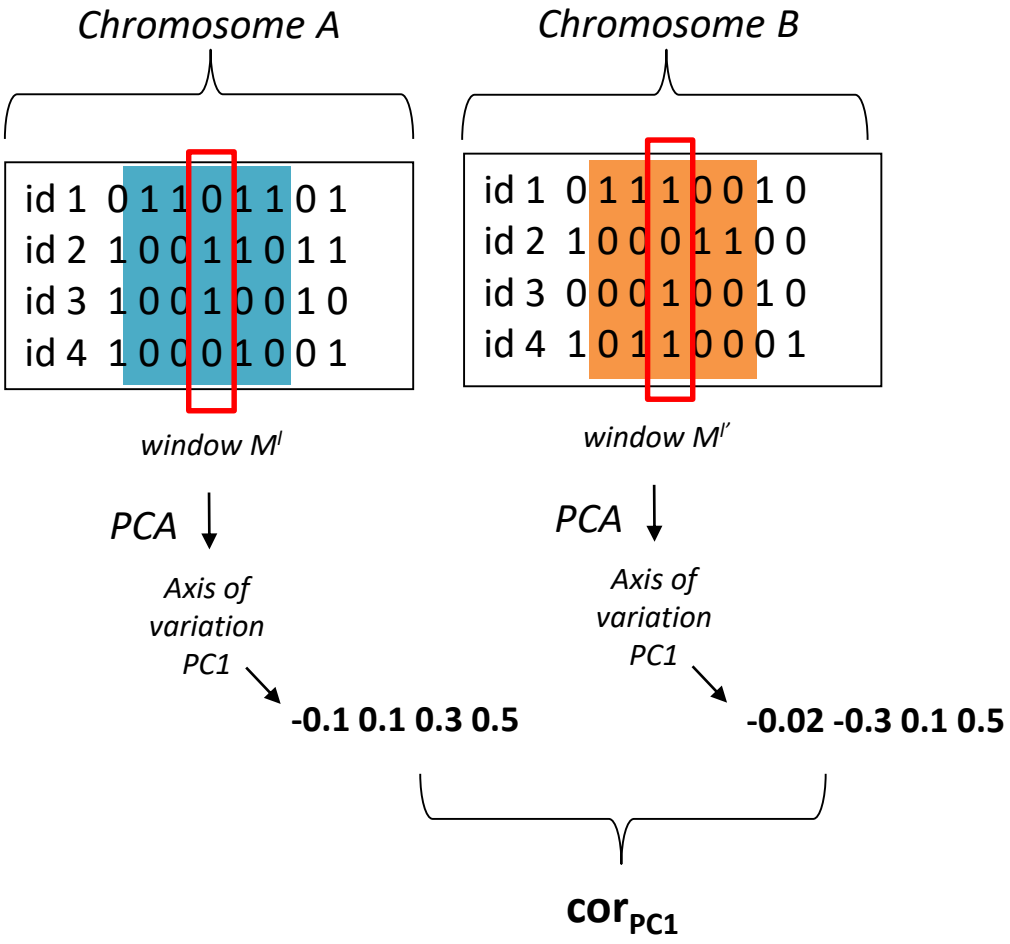
- sample has a structure
- individuals are closely related

Estimate V: Variance-Covariance matrix of sample:

$$V = \begin{pmatrix} V_{1,1} \dots V_{1,n} \dots V_{1,N} \\ \vdots \quad \vdots \quad \vdots \\ V_{n,1} \dots V_{n,n} \dots V_{n,N} \\ \vdots \quad \vdots \quad \vdots \\ V_{N,1} \dots V_{N,n} \dots V_{N,N} \end{pmatrix}$$

Linkage disequilibrium Statistics

Haplotype calculated with PCA



Correlation between two windows

$$cor_{PC1} = cor(M^{l}_{PC1}, M^{l'}_{PC1})$$

Correction by the relatedness matrix

Correlation is biased when:

- sample has a structure
- individuals are closely related

Estimate V: Variance-Covariance matrix of sample:

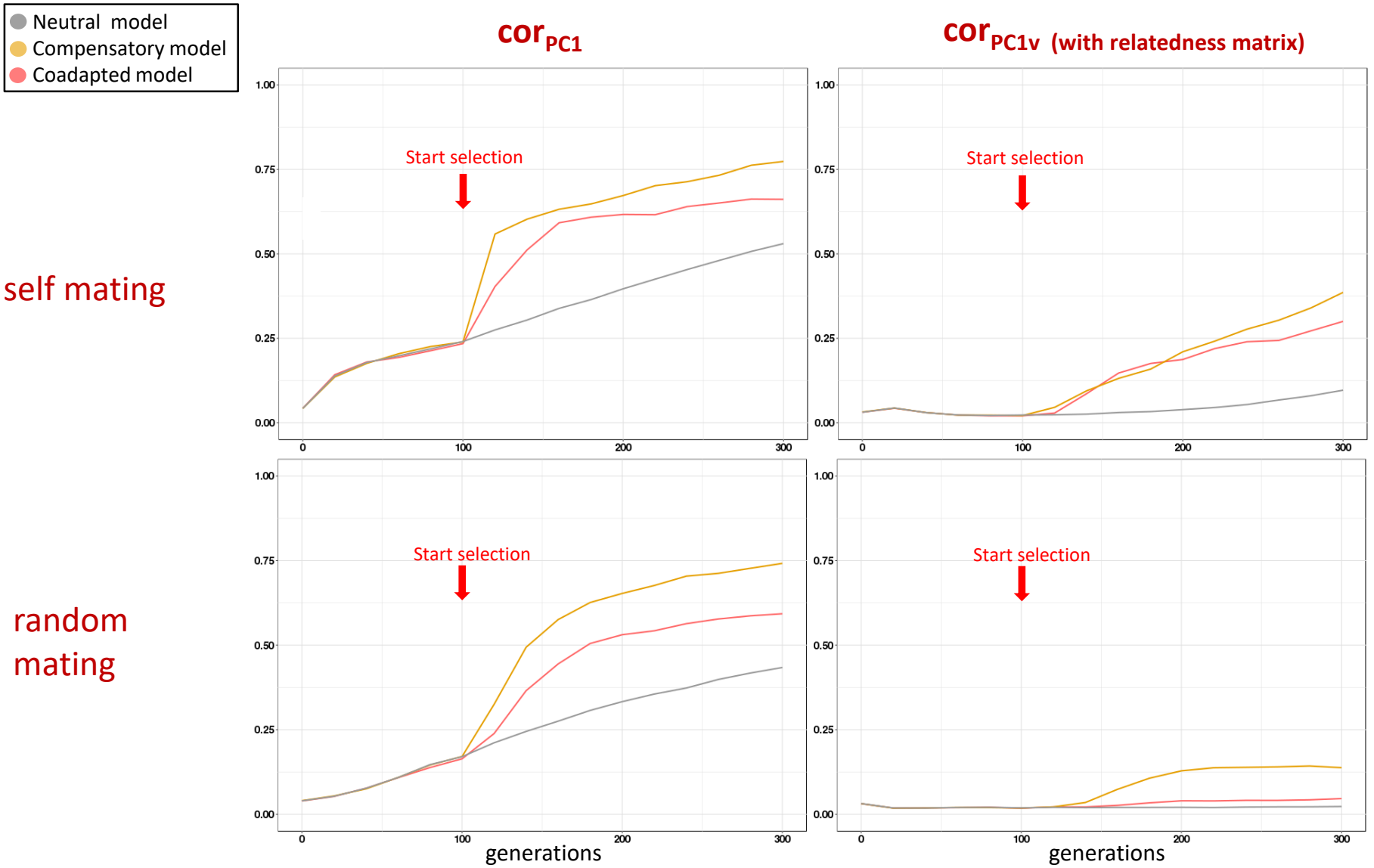
$$V = \begin{pmatrix} V_{1,1} \dots V_{1,n} \dots V_{1,N} \\ \vdots \quad \vdots \quad \vdots \\ V_{n,1} \dots V_{n,n} \dots V_{n,N} \\ \vdots \quad \vdots \quad \vdots \\ V_{N,1} \dots V_{N,n} \dots V_{N,N} \end{pmatrix}$$

Mesures to correct these biases:

$$cor_{PC1v} = cor(V^{-1/2} M^{l}_{PC1}, V^{-1/2} M^{l'}_{PC1})$$

Epistatic selection detected with linkage disequilibrium

Evolution of LD between windows and with selection into two subpopulations



Thesis: theoretical part

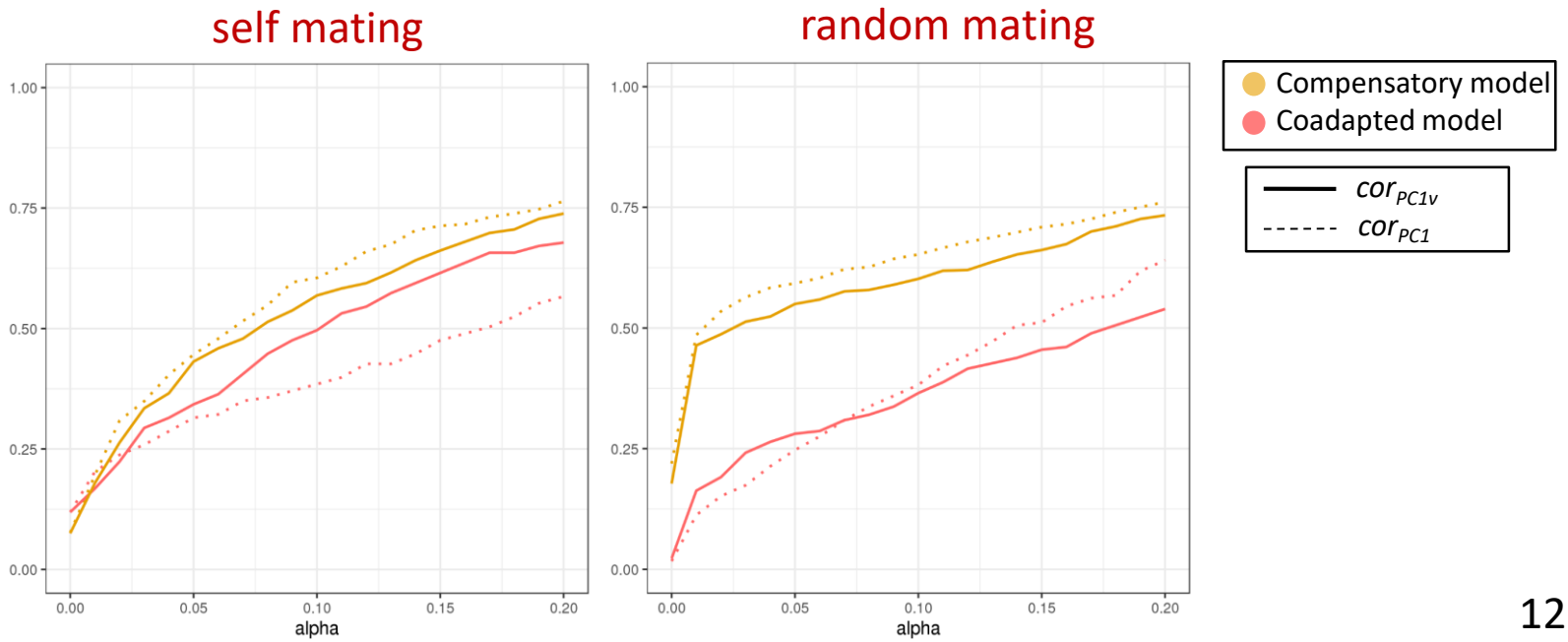
Power detection and reduction of background LD

- Distribution of test statistic *T* estimated from cor_{PC1} and cor_{PC1v} in the neutral model

$$with:\ T = \sqrt{n-2}\,\frac{cor}{\sqrt{1-cor^2}}$$

	$\tau_{(n-2)}$	Quantile 90%	Quantile 95%	Quantile 99%		Quantile 90%	Quantile 95%	Quantile 99%
	False positive	1.283254	1.647919	2.333859		1.283254	1.647919	2.333859
Self − mating	T_{corPC1}	89 %	86 %	81 %	*Random − mating*	83 %	78 %	70 %
	T_r	85 %	82 %	74 %		72 %	66 %	55 %
	T_{corPC1_v}	13 %	7 %	3 %		2.5 %	0.6 %	0.1 %
	T_{r_v}	13 %	8 %	3 %		2.8 %	0.6 %	0.2 %

- Power detection of statistics cor_{PC1} and cor_{PC1v} in coadapted and compensatory models.



Part 1 Objective: Evaluate the statistical detection of epistatic selection with a simulation approach of two independent loci.

Conclusion

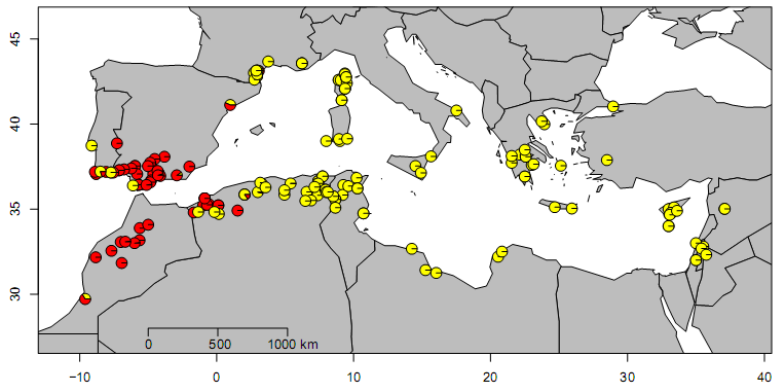
- **Evolutionary models of epistatic selection**
  - We detect fitness interaction among co-segregating variants with LD.
  - We must correct the correlation by the kinship matrix.

Part 2 Objective: Identify adaptive interactions between *Medicago truncatula* genes and identify new candidates in co-evolution with known genes.

1. Can we infer adaptive interactions between known genes
   - Do genes pairs identified in co-evolution belong to common biological pathway ?

2. Can we characterize new candidate genes or genomic regions interacting with known genes?

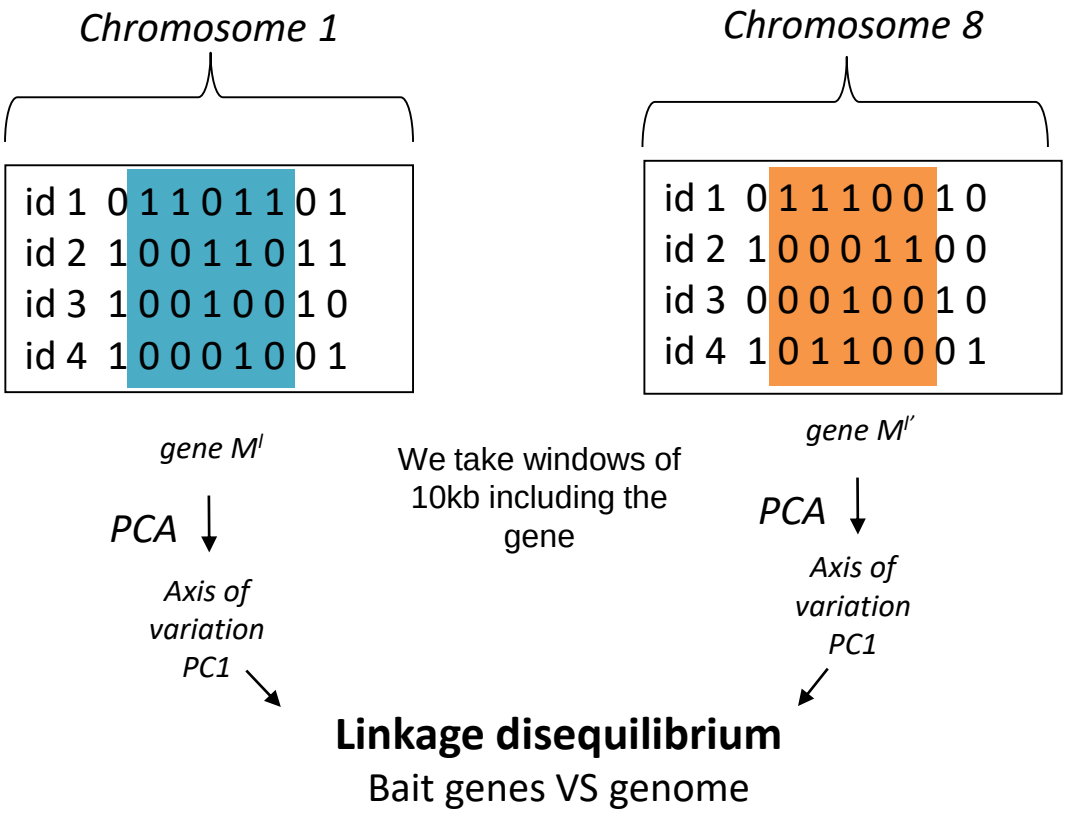
Medicago truncatula core-collection sequencing



Available data

- 22 million SNPs distributed over 8 chromosomes
- ~ 48,000 genes
- ~ 200 lines into two subpopulations (Far-West and Circum)

Linkage disequilibrium Statistics on *M. truncatula* genes



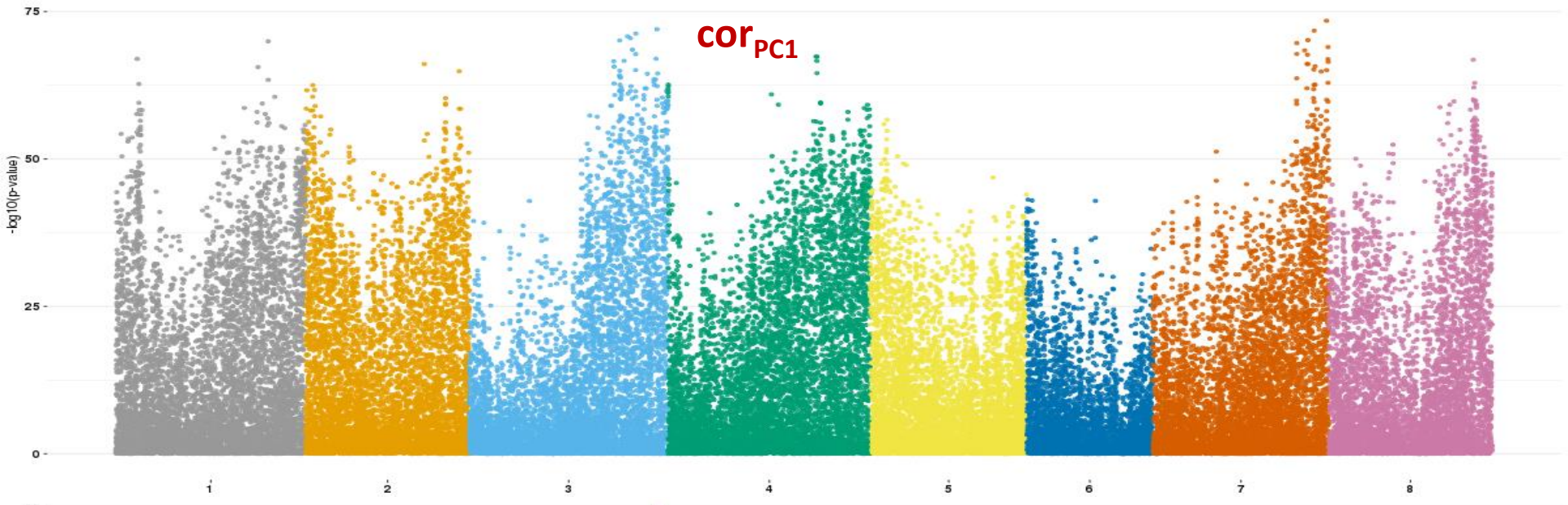
1. Matrix with 48,317 x 48,317 correlations in each subpopulations

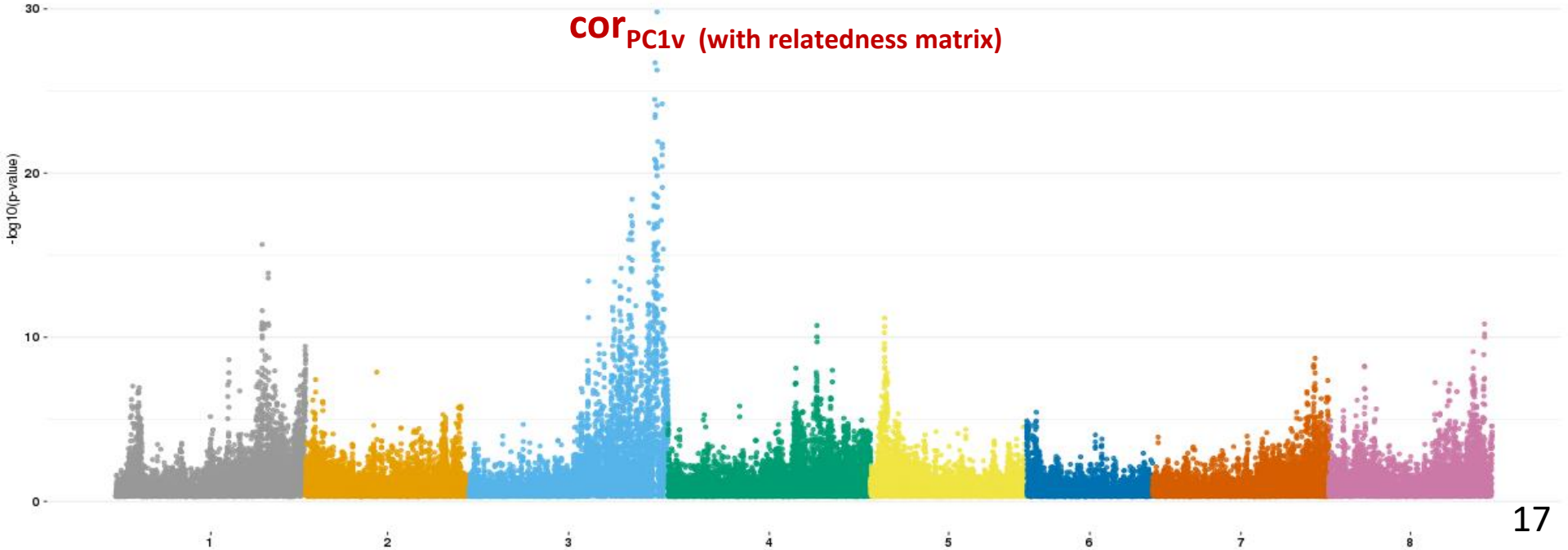
2. Statistical test on correlation (p-value)

3. Significant treshold Bonferroni (α = 5%) ➔ p-value threshold = 10^{-6}

Thesis: data analysis part

Example of MtCRA2 VS 48,000 *M. truncatula* genes

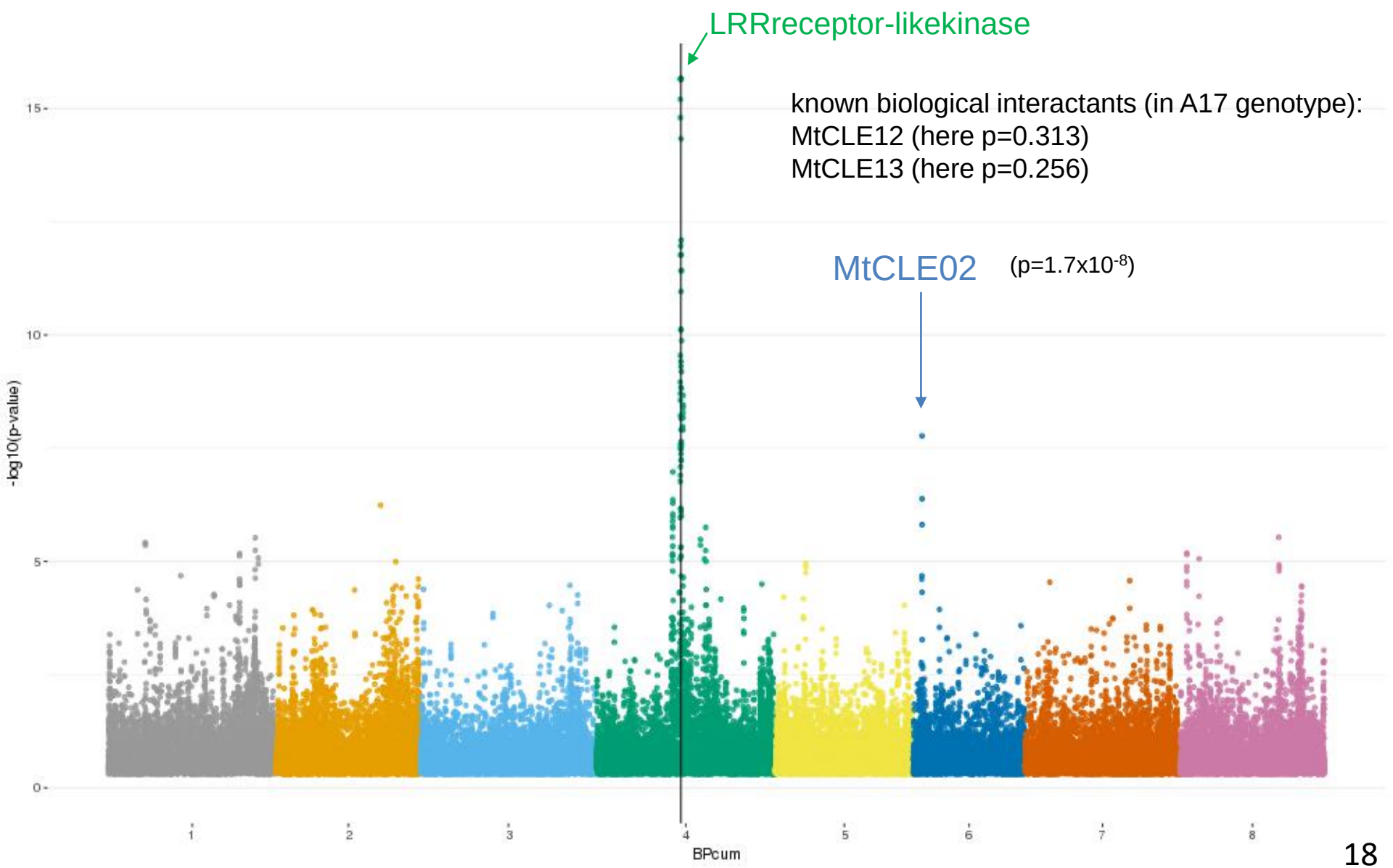




Part 2: Example of candidates bait genes

Example of MtSUNN VS 48,000 *M. truncatula* genes

cor_{PC1v} (with relatedness matrix)



Part 3 Objective: Construction of a co-evolutionary genes network with epistatic selection signatures.

1. How to interpret the network ?

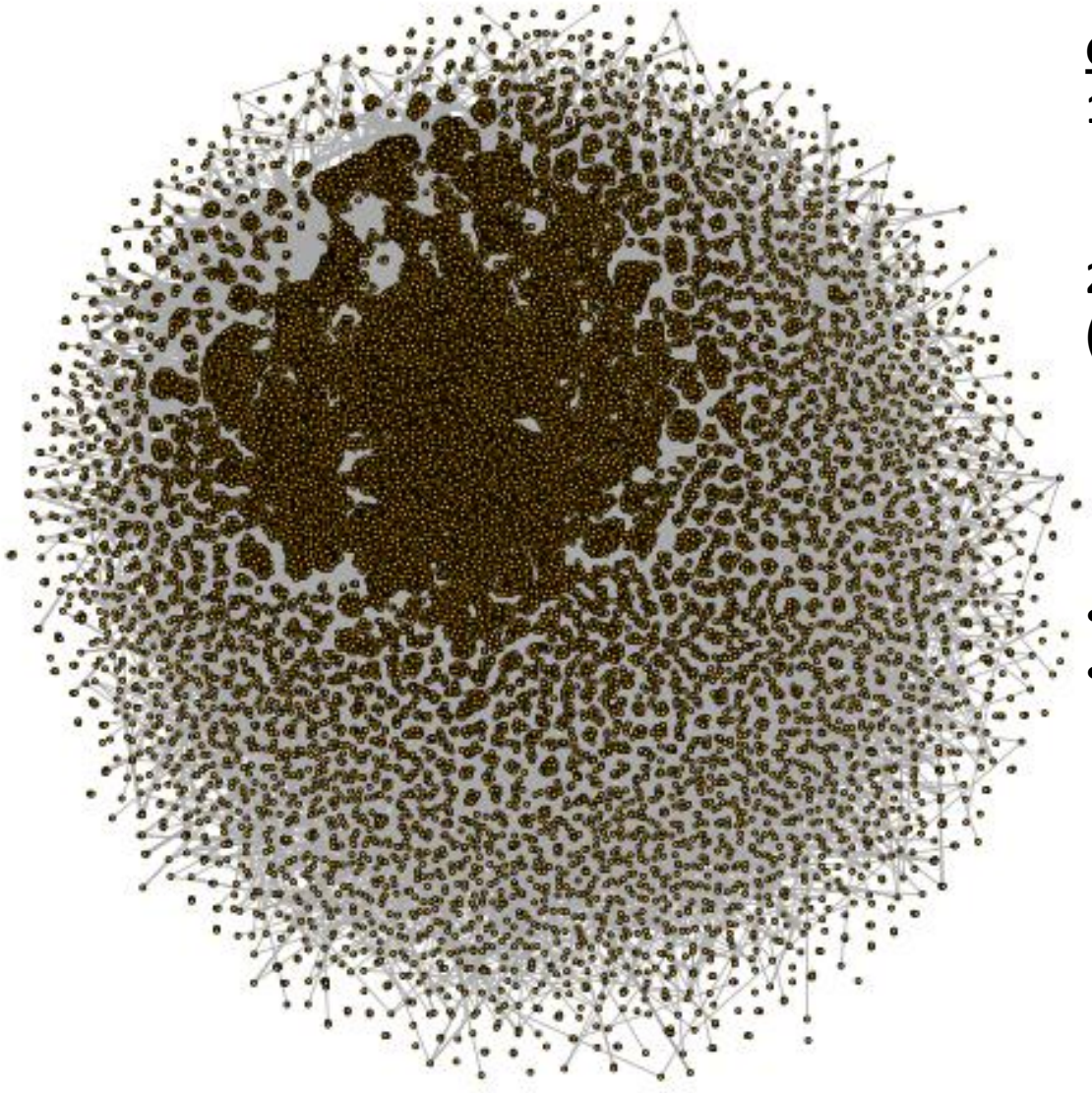
2. How to symplify the network :

    - How to filter links associated with ‘physical’ LD to ‘evolution’ LD ?

    - Can we identify epistatic selection signatures between (large) genomic regions ?

Part 3: Network construction

Genome-wide SNP interaction analyses – *M. truncatula*



Genes pairwise comparisons:

1. **Statistical test on correlation between each paire of genes.**

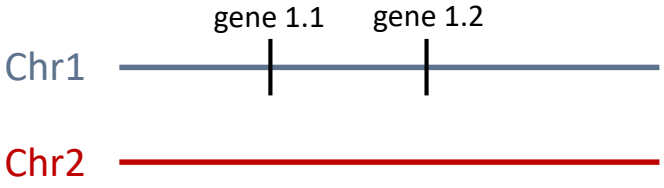
2. **Significant treshold Bonferroni (α = 5%) ➔ p-value threshold = 10^{-11}**

- **48317 nodes**
- **849715 edges**

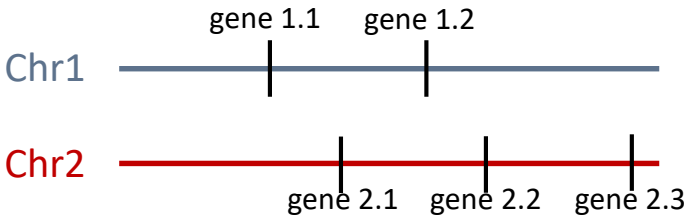
Linkage disequilibrium interpretation

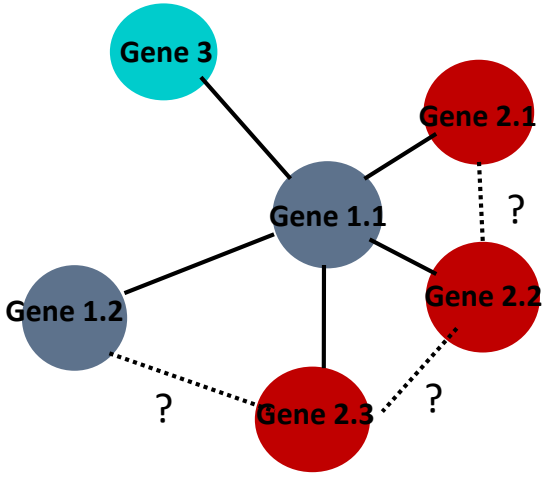
How to filter links associated with 'physical' LD to 'evolution' LD ?

Short range LD



Long range LD



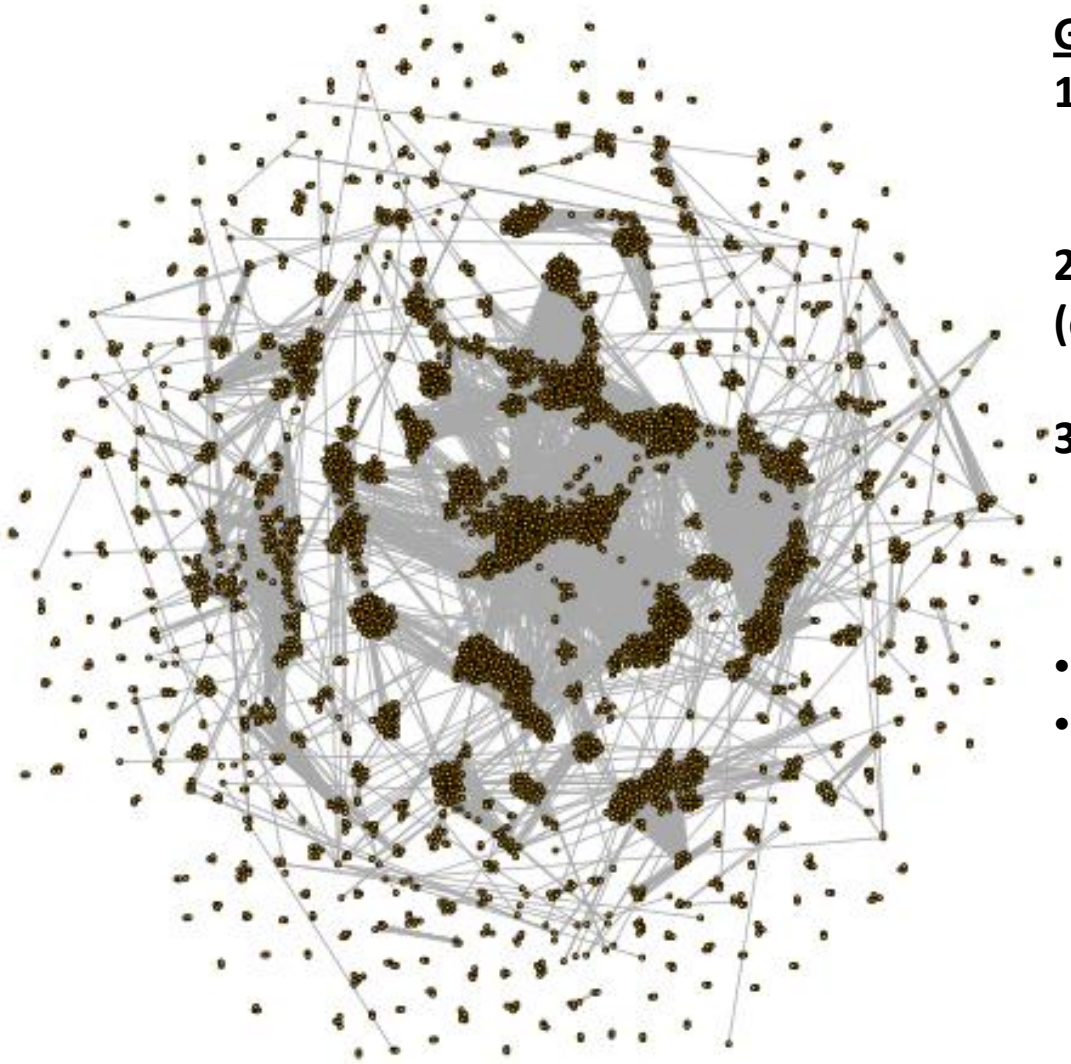


- Genes of chromosme 1 are in physical LD
- Genes of chromosome 2 are in physical LD AND they are in long range LD with gene 1.1
- Gene 3: Epistatic interaction with gene 1.1

Part 3: Network construction

Genome-wide SNP interaction analyses – *M. truncatula*

How to filter links associated with 'physical' LD to 'evolution' LD ?



Genes pairwise comparisons:

1. **Statistical test on correlation between each paire of genes**

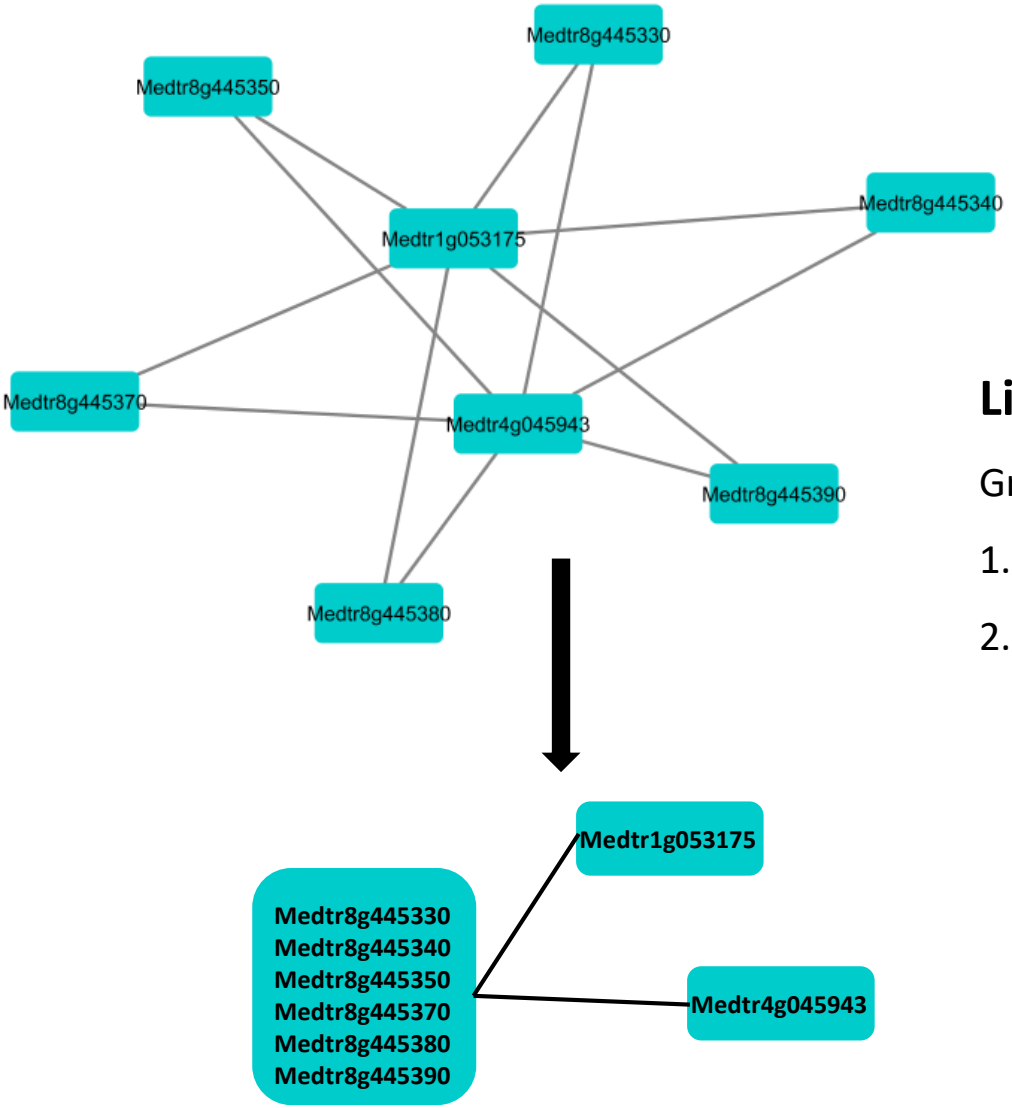
2. **Significant treshold Bonferroni (α = 5%) ➔ p-value threshold = 10^{-11}**

3. **Only gene pairs on different chromosomes**

- **4568 nodes**
- **42114 edges**

Genome-wide SNP interaction analyses – *M. truncatula*

Identify epistatic selection signatures between (large) genomic regions



Linkage disequilibrium interpretation

Group genes into super nodes:

1. Criteria: Distance and LD between gene pairs
2. Representation and analysis of the interactions between island:
   - How many island ?
   - Intensity of interaction between island ?

Objective: Development of genetic methods to detect adaptive interactions between genes in *Medicago truncatula*

Conclusion

- **Part 2: Detection of genes under epistatic selection in *Medicago truncatula* and in humans.**
  - SNP analyses in *Medicago truncatula:* "bait" methods. -> Results analysis in progress

- **Part 3: Detection of coadapted clusters by genes correlation network analysis**
  - Generate resource database containing genes pairwise correlation in *M. truncatula* genome.
  - **Perspective**
    - Identify epistatic selection between genomic windows.
    - Join these results to functional annotations, example: symbiotic island.

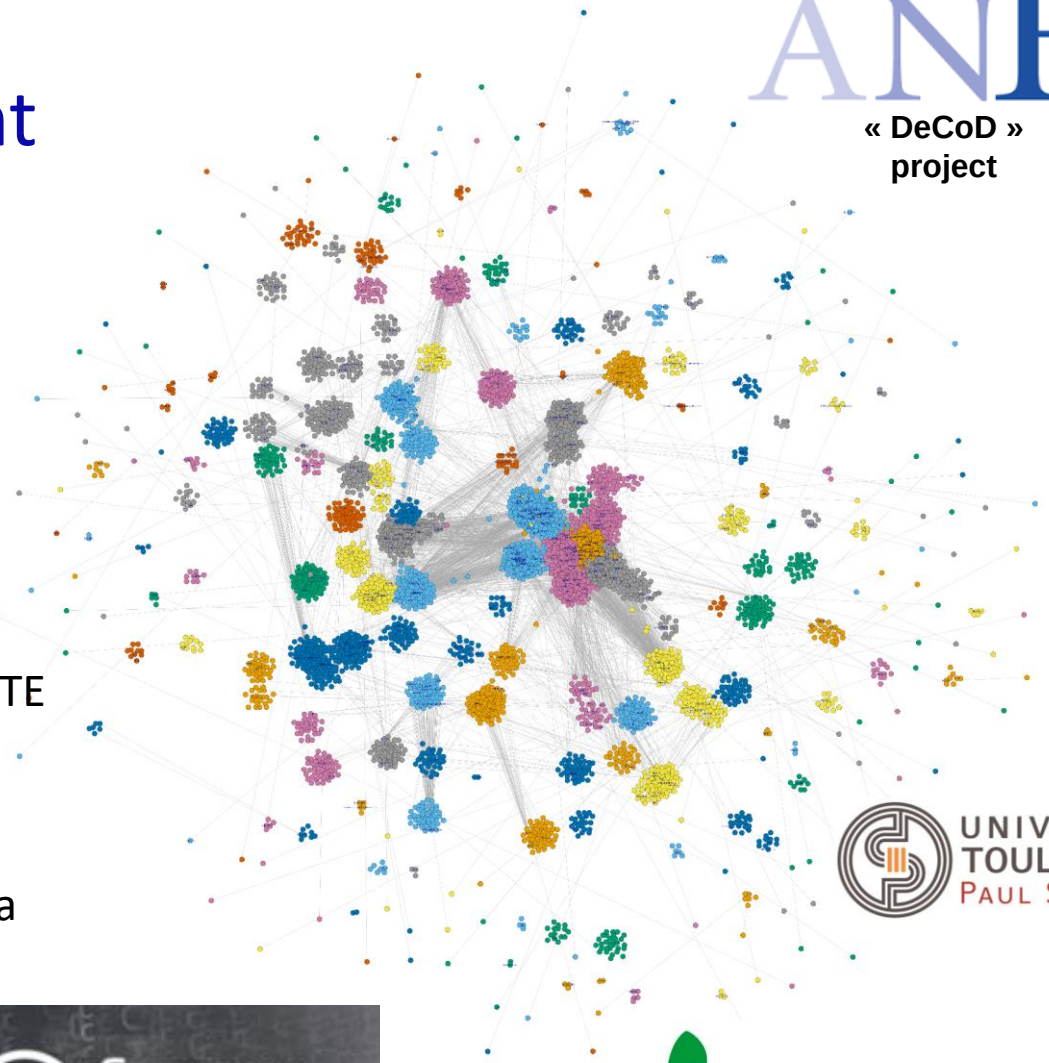
Thank you for your attention

Acknowledgment

AGENCE NATIONALE DE LA RECHERCHE ANR

« DeCoD » project

cnrs



LRSV – IPM team:
Maxime Bonhomme
Christophe Jacquet
LRSV – MYCO team:
Pierre-Marc Delaux

IPS2:
Marie Laure MARTIN MAGNIETTE

INR Toulouse:- Brigitte Mangin
INRA Nancy – Stéphane De Mita

UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

GenoToul Bioinfo

RSV Laboratoire de Recherche en Sciences Végétales