Identification of disease susceptibility loci using a co-evolution measure and haplotype phylogenies

Claire Bardel, Pierre Darlu and Emmanuelle Génin

INSERM U535, Villejuif, France

Introduction

Background

- Developpement of molecular genetics ⇒ lots of markers available within genes, in particular SNPs
- Haplotypic methods allow to use the joint information of several markers to test for association between a gene and a disease:
  - but, increase in the number of markers → increased number of haplotypes → low power of the association test
- Several studies have proposed to group haplotypes according to their evolutionary history to perform association tests (Templeton *et al.*, 1987; Seltman *et al.*, 2003; Durrant *et al.*, 2004, Bardel *et al.* 2005)
- The evolutionary history of haplotypes, represented by a phylogenetic tree can also provide information about the localization of disease susceptibility (DS) SNPs

⇒ Presentation of a method to look for susceptibility loci, study of its efficiency in various conditions and application to data

Method

H_7: 110...

H_6: 000...

H_1: 000...

H_5: 001...

H_4: 011...

H_3: 101...

H_2: 101...

Initial data

- 7 haplotypes formed by several SNPs
- Focus on 3 particular SNPs

Method

cc : number of cases/controls

cc: 12/6
H_7: 110...

H_6: 000...
cc: 0/3

H_1: 000...
cc: 4/9

H_5: 001...
cc: 8/18

H_4: 011...
cc: 2/9

H_3: 101...
cc: 24/10

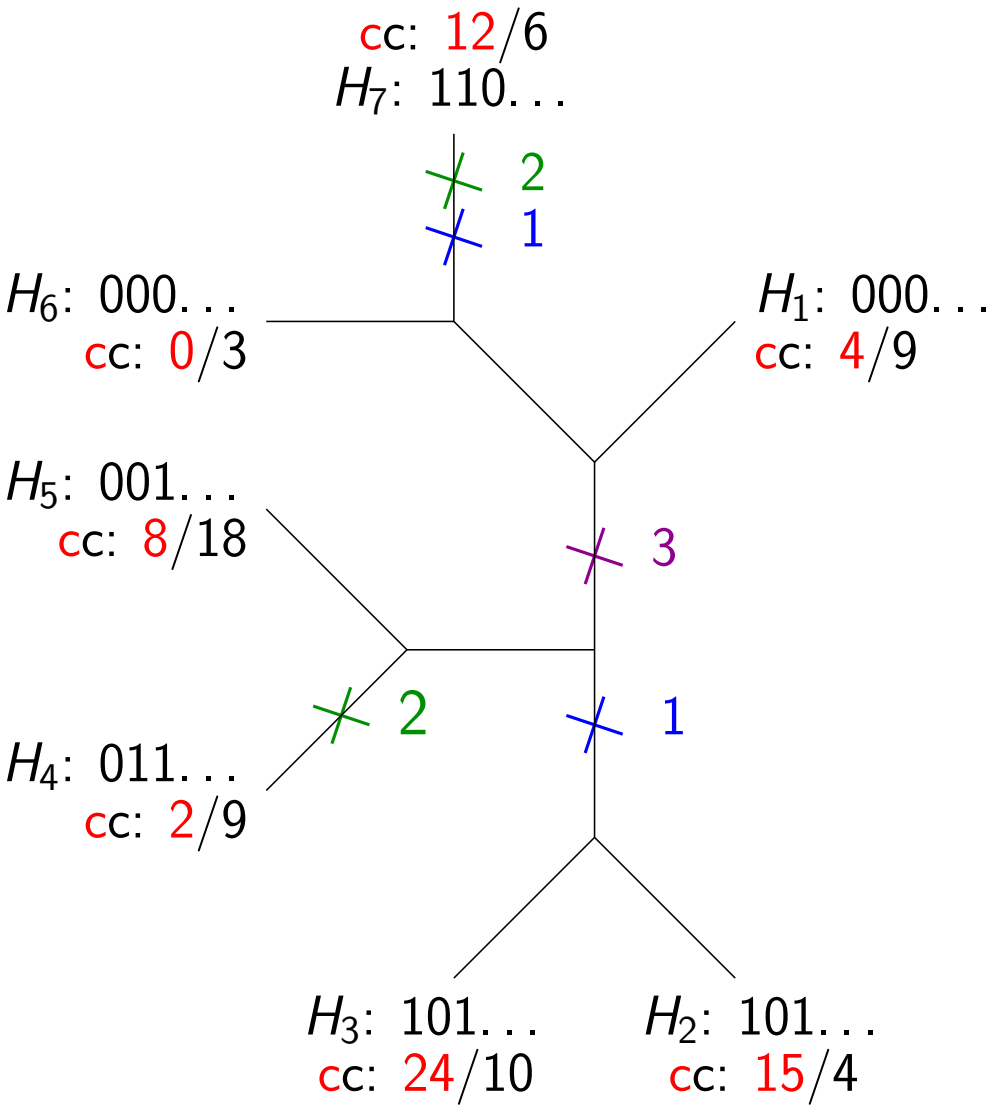
H_2: 101...
cc: 15/4

Initial data

- 7 haplotypes formed by several SNPs
- Focus on 3 particular SNPs
- Haplotypes carried by cases and controls

Method

cc : number of cases/controls

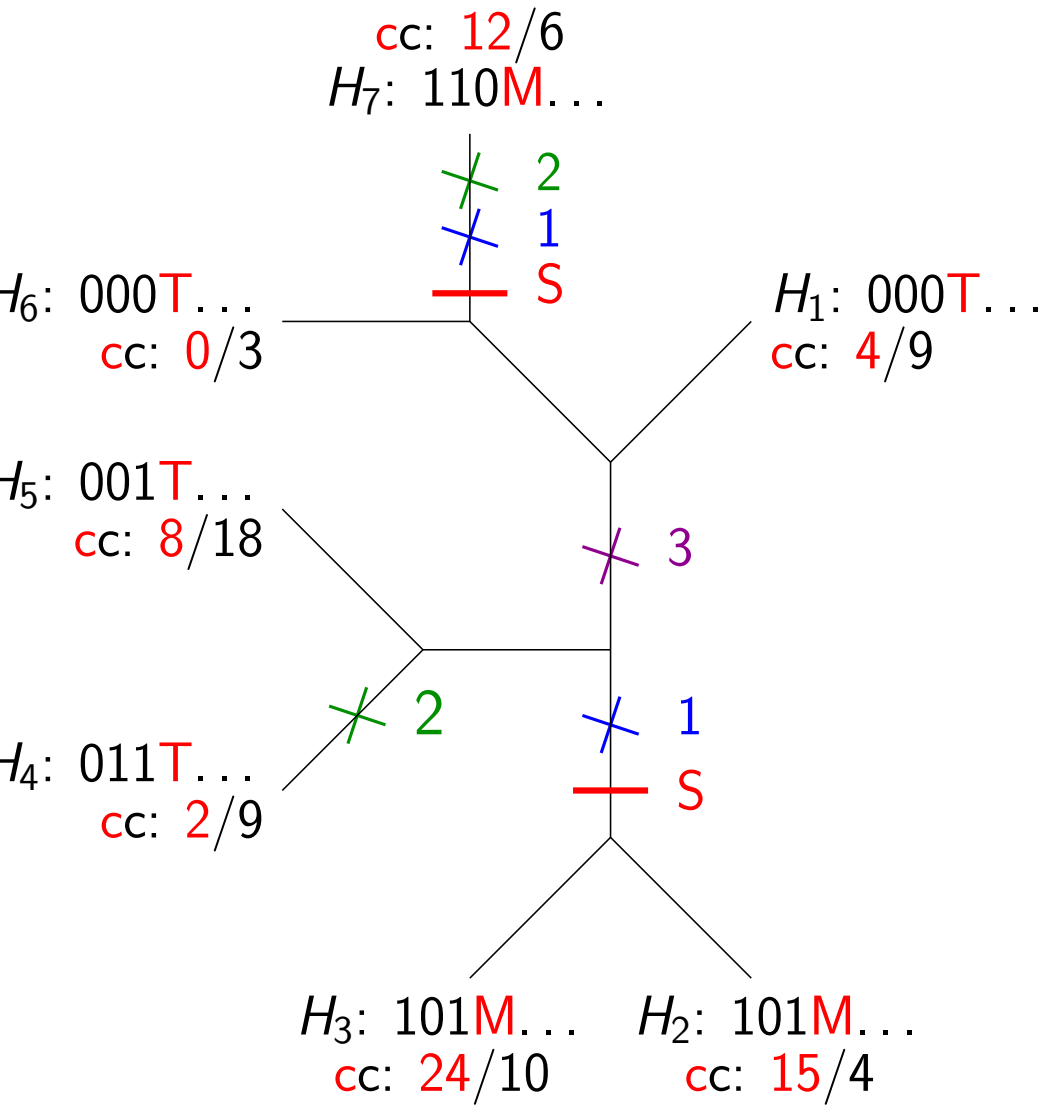


1/ Construction of the haplotype tree

- Use of a parsimony method (software PAUP)
  - minimization of the number of evolutionary steps (i.e. $0 \longrightarrow 1$ or $1 \longrightarrow 0$ changes on the tree)
- Identification of the SNPs state changes on the tree branches

Method

cc : number of cases/controls

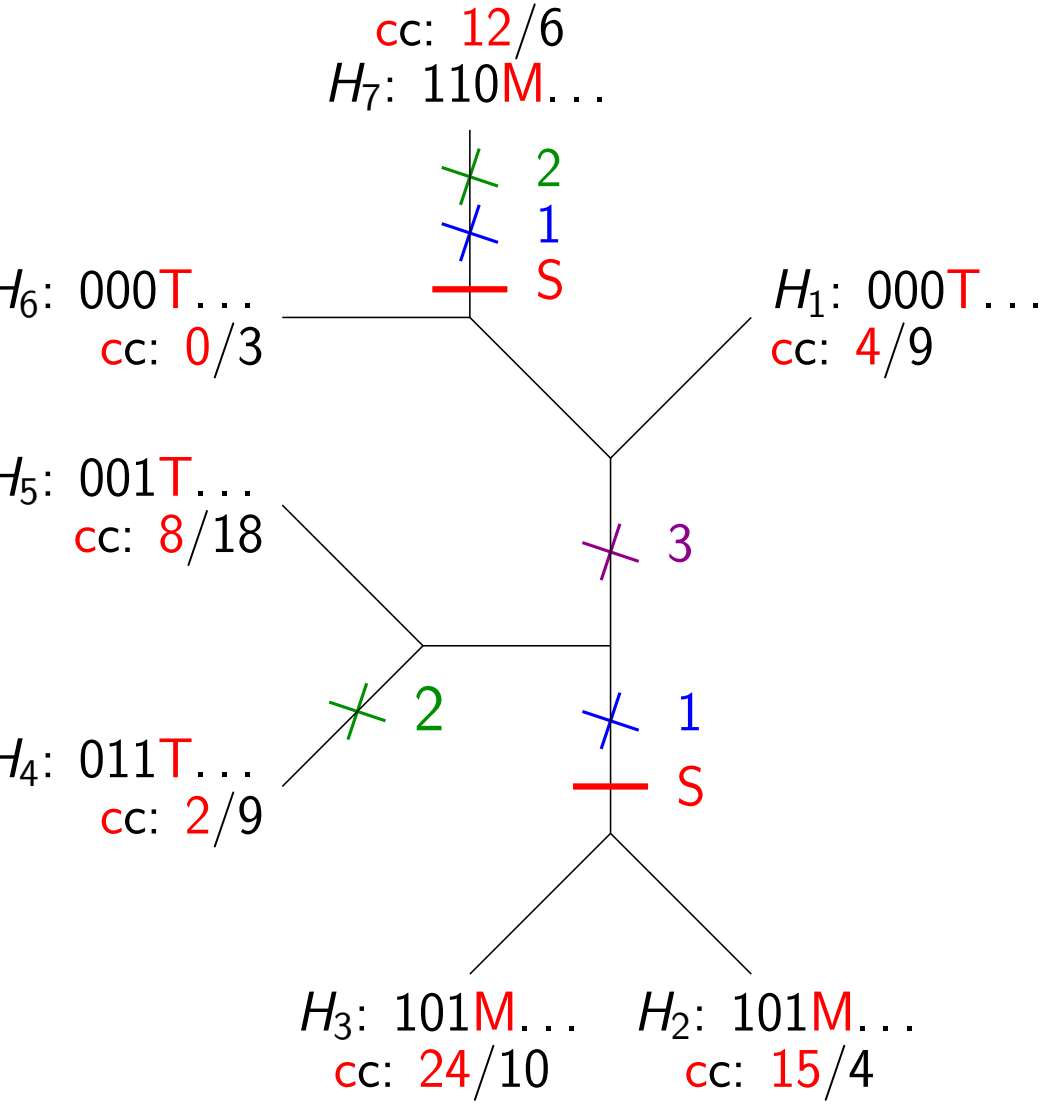


2/ Definition of a new character (S)

- S is defined for each haplotype
- S has 2 states: M and T
- The state M (resp.T) is attributed to haplotypes carried by a significant majority of cases (resp. controls)
- Identification of S state changes on the tree

Method

cc : number of cases/controls



3/ Definition of a co-evolution index : V_i

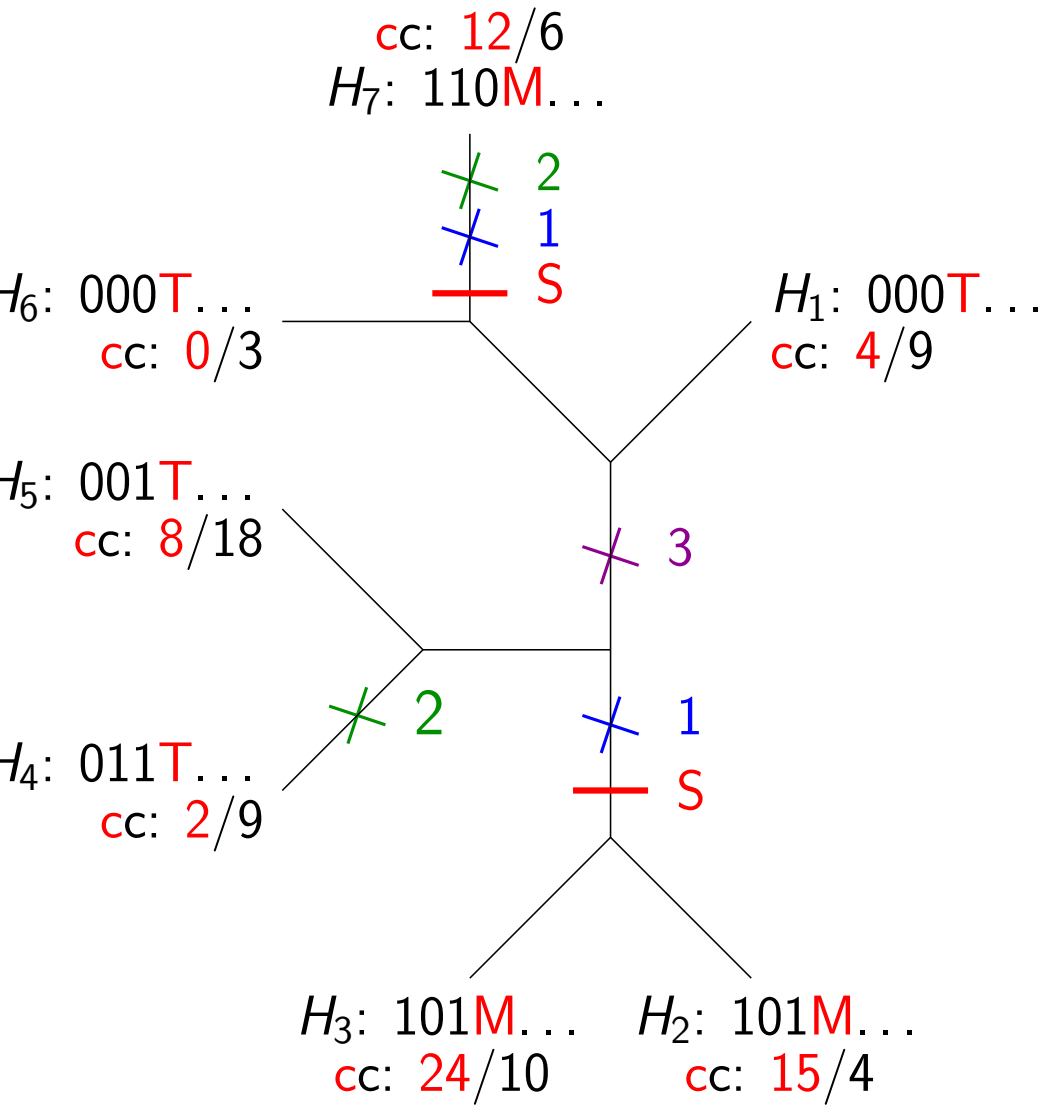
- Count the number O_i of observed co-mutations for SNP i

  $O_1 = 2$; $O_2 = 1$; $O_3 = 0$
- Compute the number E_i of expected co-mutations under the hypothesis of random distribution of mutations on the 11 branches of the tree
  - $E_1 = \frac{2}{11} \times \frac{2}{11} \times 11 = 0.4$
  - $E_2 = 0.4$; $E_3 = 0.2$
- Definition of V_i:

$$V_i = \frac{O_i - E_i}{\sqrt{(E_i)}}$$

Method

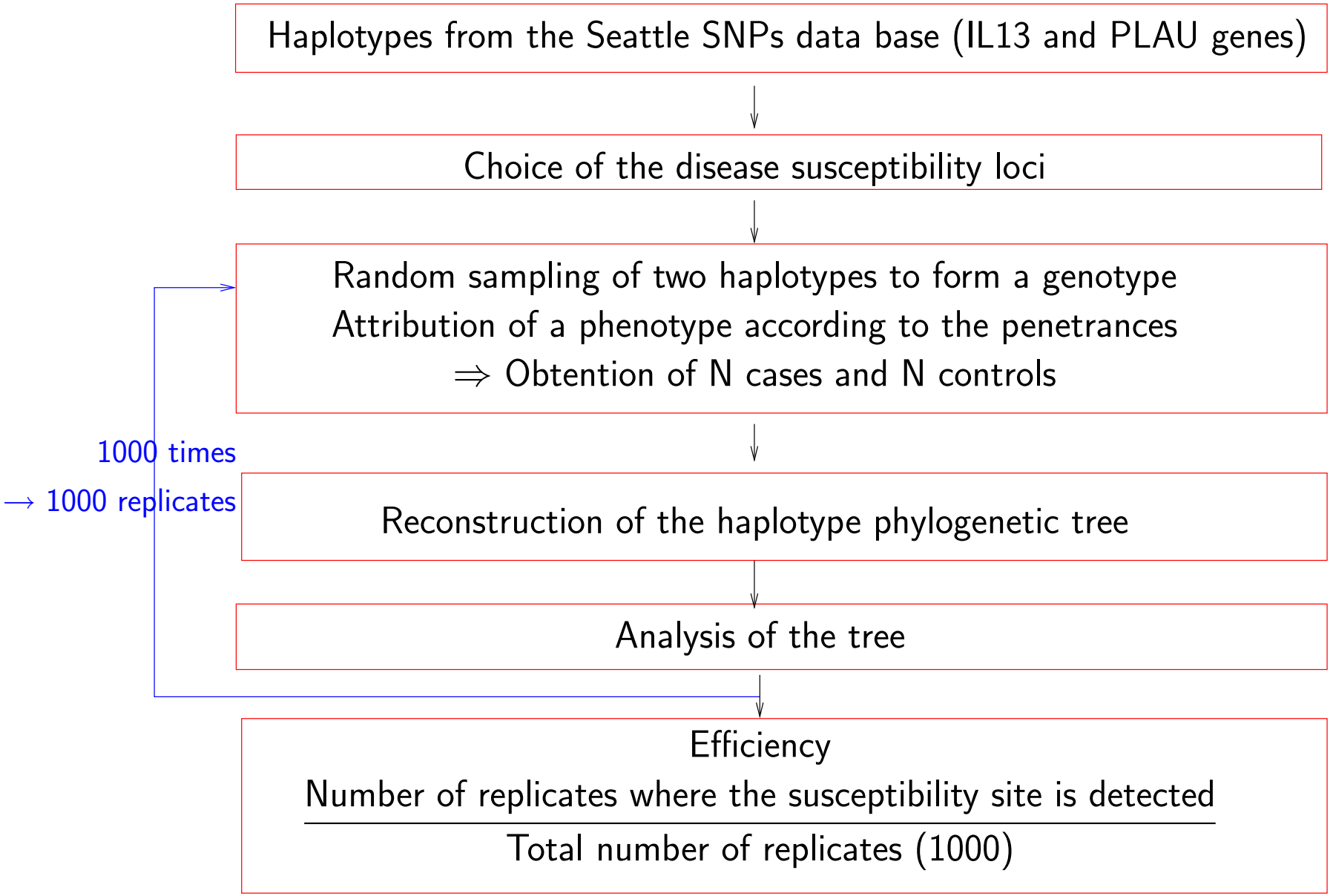
cc : number of cases/controls



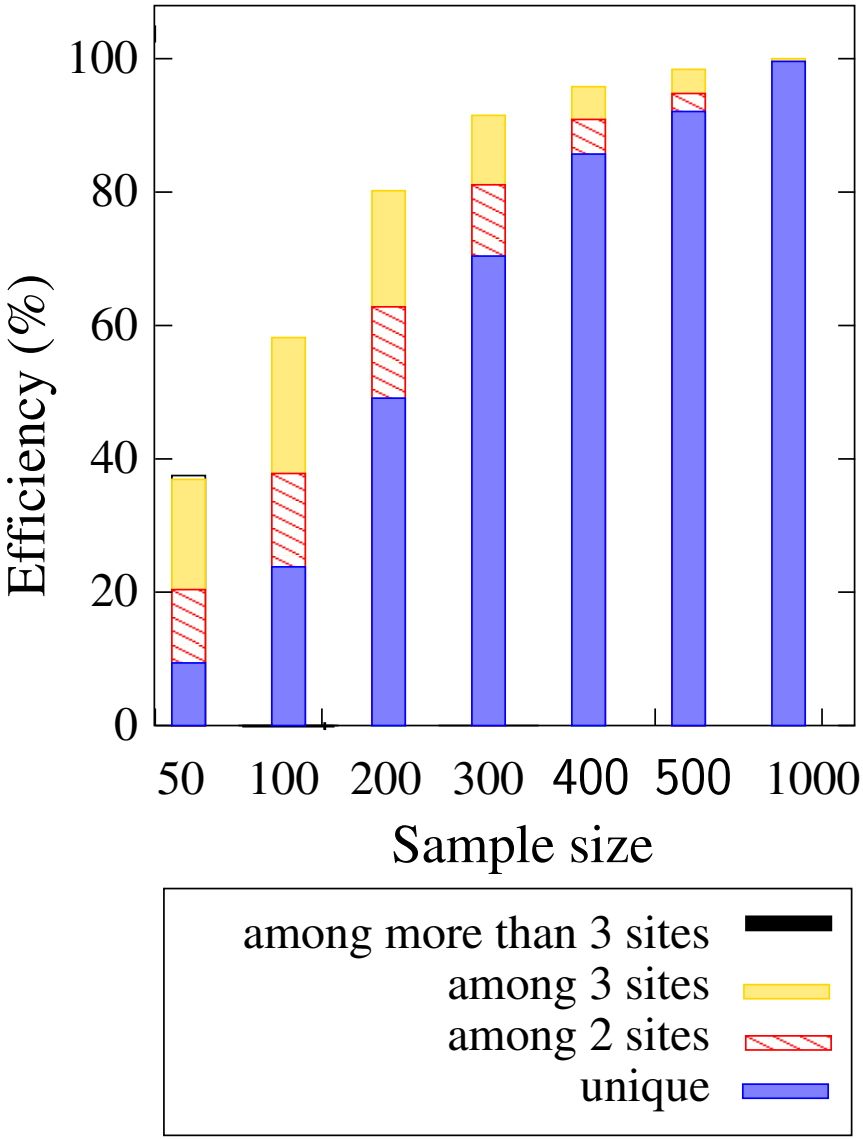
4/ Identification of the disease susceptibility (DS) site

- Compute the V_i for each SNP
  - $V_1 = 2.7$ $V_2 = 1.1$ $V_3 = -0.4$
- If there are several equiparsimonious trees, foreach site, sum the V_i over all the trees
- The sites with the highest V_i are putative DS sites

Simulation process



Results (I)



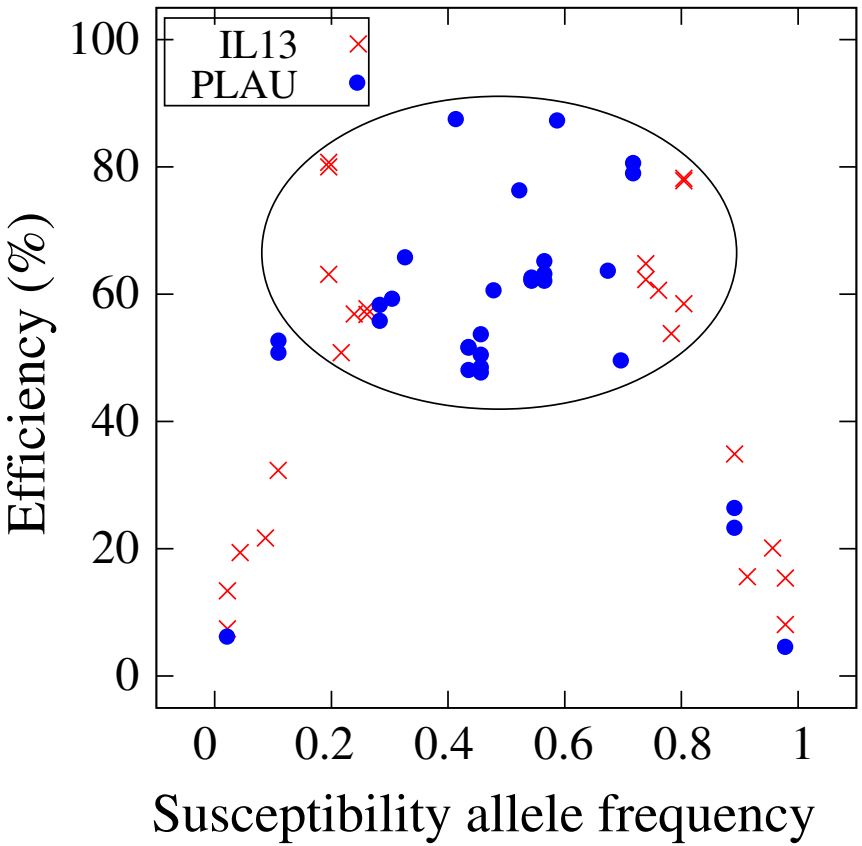
Influence of the sample size

- As expected, the efficiency increases with the sample size
- When the sample size increases, the proportion of replicates in which the site is the only one detected also increases

Simulated disease model:

- freq of the susc. allele: 0.3
- GRR heterozygote: 2
- GRR homozygote: 10

Results (II)



sample size: 200 cases + 200 controls
penetrance: 0.03 (hom non carrier)
0.06 (heterozygotes
0.30 (hom carrier)

Influence of frequency f of the DS allele

In our simulation conditions,

- efficiency is maximum if $0.2 < f < 0.8$ (complex diseases)
  - efficiency is not correlated with f
- if $f < 0.2$, or $f > 0.8$, the efficiency is low and correlated with f
  - role of our penetrances (high rate of phenocopies)

Application to data on the DRD2 gene and schizophrenia

The data set (Dubertret *et al.*, 2004)

- 103 trios (parents + 1 affected child)
  - transmitted haplotypes $\longrightarrow$ case sample
  - non transmitted haplotypes $\longrightarrow$ control sample
- 7 SNPs + 1 microsatellite are genotyped in the DRD2 gene and the X-kinase gene
- Haplotypes reconstructed with `FAMHAP` v1.5 $\longrightarrow$ 59 $\neq$ haplotypes
- 1000 equiparsimonious trees reconstructed with `PAUP`

Results

- The V_i are calculated for the 7 SNPs but not for the microsatellite
- The site with the highest V_i is **SNP 3 (TaqI A1/A2)**, A2 being the susceptibility allele
- This SNP has already been reported to be involved in the determinism of schizophrenia (Dubertret *et al.*, 2001, 2004)

Conclusion and perspectives

Conclusion

- Development of a new method to localize DS loci
  Implementation in the software `ALTree` (available soon)
- Study of its efficiency:
  - It highly depends on the sample size
  - It is more powerful when f is between 0.2 and 0.8
  - Other simulations suggest that it also depends on the penetrances and on the position of the DS site changes on the tree
  - Comparison with other methods $\longrightarrow$ it can improve the localisation of the DS site, especially when more than 1 DS site is simulated

Future work

- Study the impact of the phylogenetic reconstruction method:
  ML $\longrightarrow$ attribution of a weight to character state changes
- Study of the impact of haplotype reconstruction on the method and possibly take the uncertainty in the haplotype reconstruction into account

How to define the state of the character S?

Attribution of a new character S to each haplotype corresponding to the disease status of the haplotype

For each haplotype, the state of the character S is:

- T (control) if $p_h < p_0 - \sqrt{\frac{p_h \times (1-p_h)}{n_h}}$
- M (case) if $p_h > p_0 + \sqrt{\frac{p_h \times (1-p_h)}{n_h}}$
- ? (unknown) else

Where:

- p_h = proportion of cases among the carrier of haplotype h
- p_0 = proportion of cases in the whole sample of haplotypes
- n_h = number of individuals carrying haplotype h

Exact computation of V_i

Definition of V_i

- **Definition:** On a tree t, for a given site i and a given transition (e.g. $0 \longrightarrow 1$), $V_i^{0\to1}$ measures the co-evolution between transition $0 \longrightarrow 1$ of site i and the character S.
- **Computation:**
  - $E_i^{0\to1}$: number of expected co-mutations of S and i:

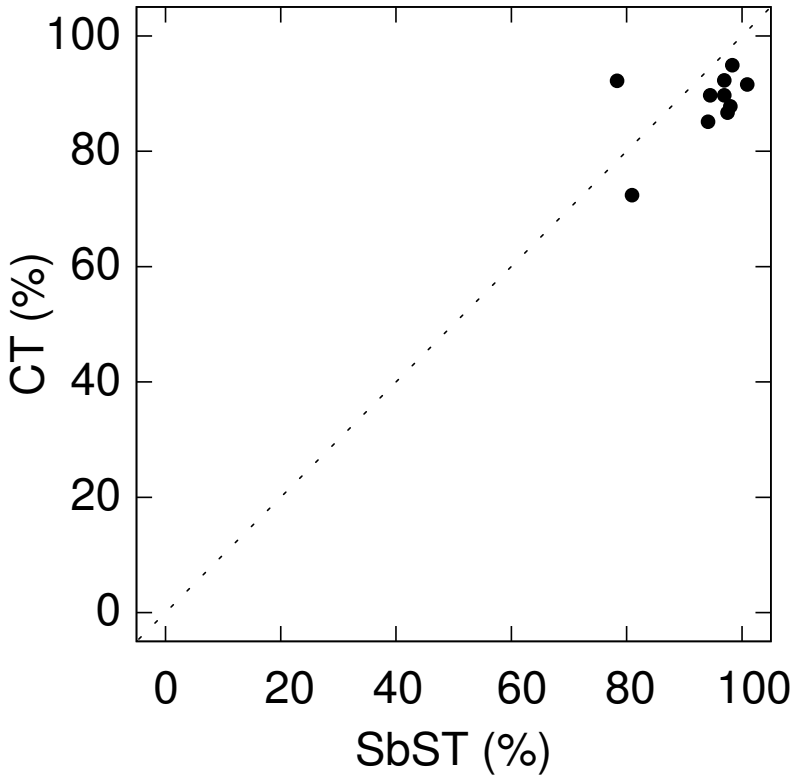
$$E_i^{0\to1} = \frac{(m_i^{0\to1} \times s^{T\to M}) + (m_i^{1\to0} \times s^{M\to T})}{b}$$

$m_i^{0\to1}$: nb transitions $0 \longrightarrow 1$ of i
$s^{T\to M}$: nb transitions $T \longrightarrow M$ of S o
b : nb branches of tree t

  - $R_i^{0\to1}$: number of observed co-mutations of S and i

$$\Rightarrow \quad \boxed{\begin{cases} V_i^{0\to1} = 0 & \text{if } E_i^{0\to1} = 0 \\ V_i^{0\to1} = \dfrac{R_i^{0\to1} - E_i^{0\to1}}{\sqrt{E_i^{0\to1}}} & \text{if } E_i^{0\to1} \neq 0 \end{cases}}$$

Comparison with a locus by locus method



D is the at risk allele

frequency: $f(D) = 0.2$

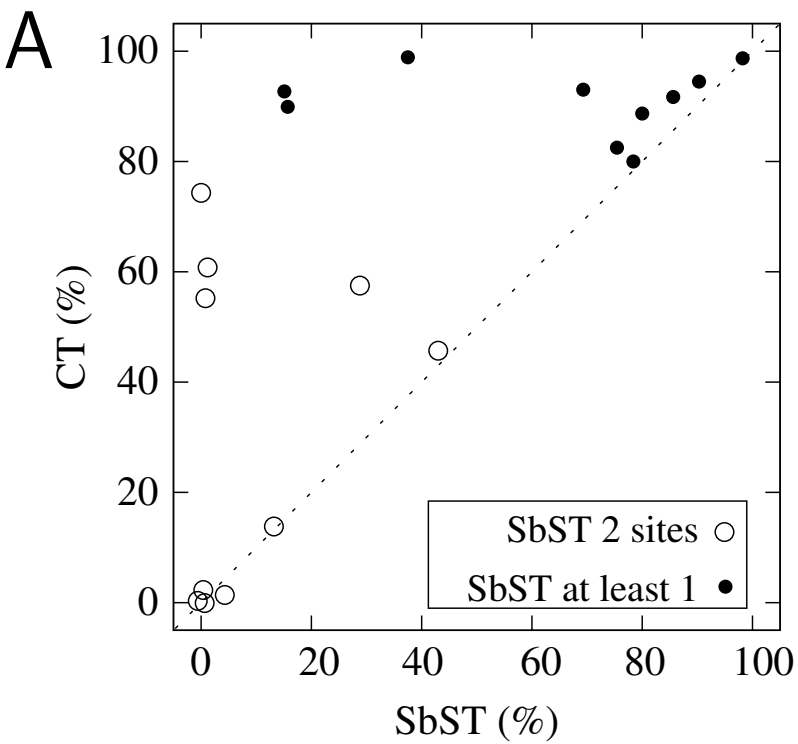
penetrance:

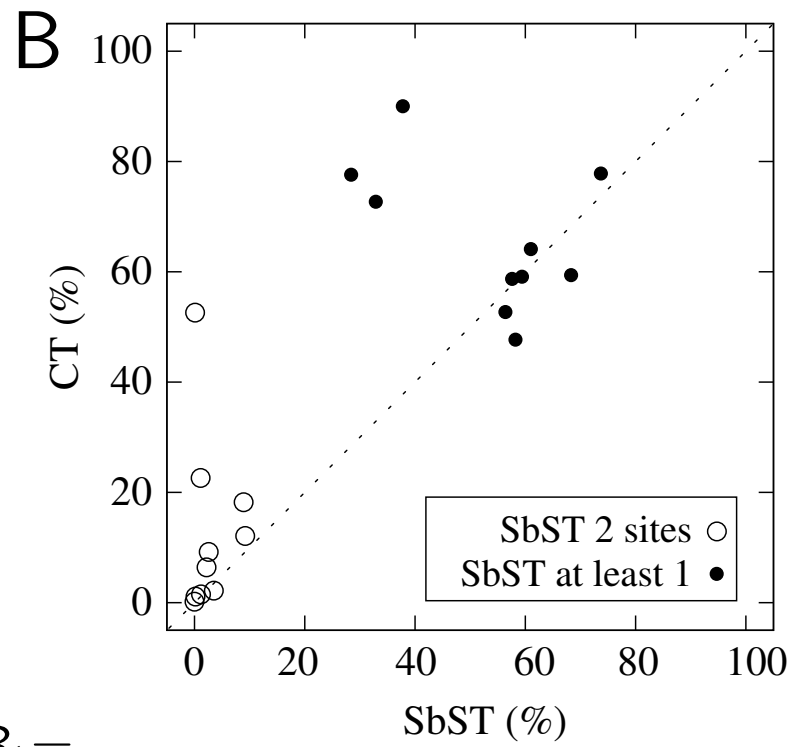
p(dd)=0.03

p(Dd)=0.06

p(DD)=0.3

Comparison with a locus by locus method (II)





$A_1B_1 =$ A_1B_1

at risk haplotype
frequency: $f(A_1B_1) = 0.25$
penetrances: $p(A_1B_1) = 0.9$
other penetrances $p = 0.3$

at risk haplotype
frequency: $f(A_1B_1) = 0.25$
penetrances: $p(A_1B_1) = 0.6$
other penetrances $p = 0.3$