Utilisation de phylogénies d'haplotypes pour identifier des locus à risque

Claire Bardel[1], Pierre Darlu[2] et Emmanuelle Génin[2]

1: UMR 5145, CNRS MNHN Univ. Paris VII, Paris
2: INSERM U535, Villejuif

Introduction (1)

Contexte général : étude des maladies complexes

- Dues à des facteurs génétiques et/ou environnementaux
  - ex : maladies cardio-vasculaires, cancers, maladies psychiatriques, ...

Stratégies de recherche de gènes impliqués dans le déterminisme des maladies

- études d'association : comparaison de l'information génétique d'échantillons de malades et de témoins
- Stratégies genome-wide / Stratégies gène candidat

Introduction (2)

Les méthodes haplotypiques basées sur des phylogénies

- Développement des techniques de biologie moléculaire
  $\Rightarrow$ nombreux marqueurs disponibles au sein de gènes
- Méthodes haplotypiques : utilisation de l'information conjointe de plusieurs marqueurs pour rechercher une association entre un gène et une maladie
  - Mais : $\nearrow$ nombre de marqueurs $\Rightarrow$ $\nearrow$ nombre d'haplotypes
    $\Rightarrow$ faible puissance du test
- Plusieurs tests d'association basés sur le regroupement des haplotypes selon leur histoire évolutive ont été proposés (Templeton *et al.*, 1987 ; Seltman *et al.*, 2003 ; Durrant *et al.*, 2004, Bardel *et al.* 2005)
- La phylogénie des haplotypes peut aussi fournir des informations concernant la localisation des locus à risque pour la maladie

$\Rightarrow$ Présentation d'une méthode d'identification des locus à risque, étude de son efficacité et application à des données

Méthode

$H_7 : 110 \quad \ldots$

$H_6 : 000 \quad \ldots$

$H_1 : 000 \quad \ldots$

$H_5 : 001 \quad \ldots$

$H_4 : 011 \quad \ldots$

$H_3 : 101 \quad \ldots$

$H_2 : 101 \quad \ldots$

Données initiales

- 7 haplotypes formés par plusieurs SNPs
- Illustration sur 3 des SNPs

Méthode

mt : nombre de malades/témoin

mt : 12/6
H_7 : 110 ...

H_6 : 000 ...
mt : 0/3

H_1 : 000 ...
mt : 4/9

H_5 : 001 ...
mt : 8/18

H_4 : 011 ...
mt : 2/9

H_3 : 101 ...
mt : 24/10

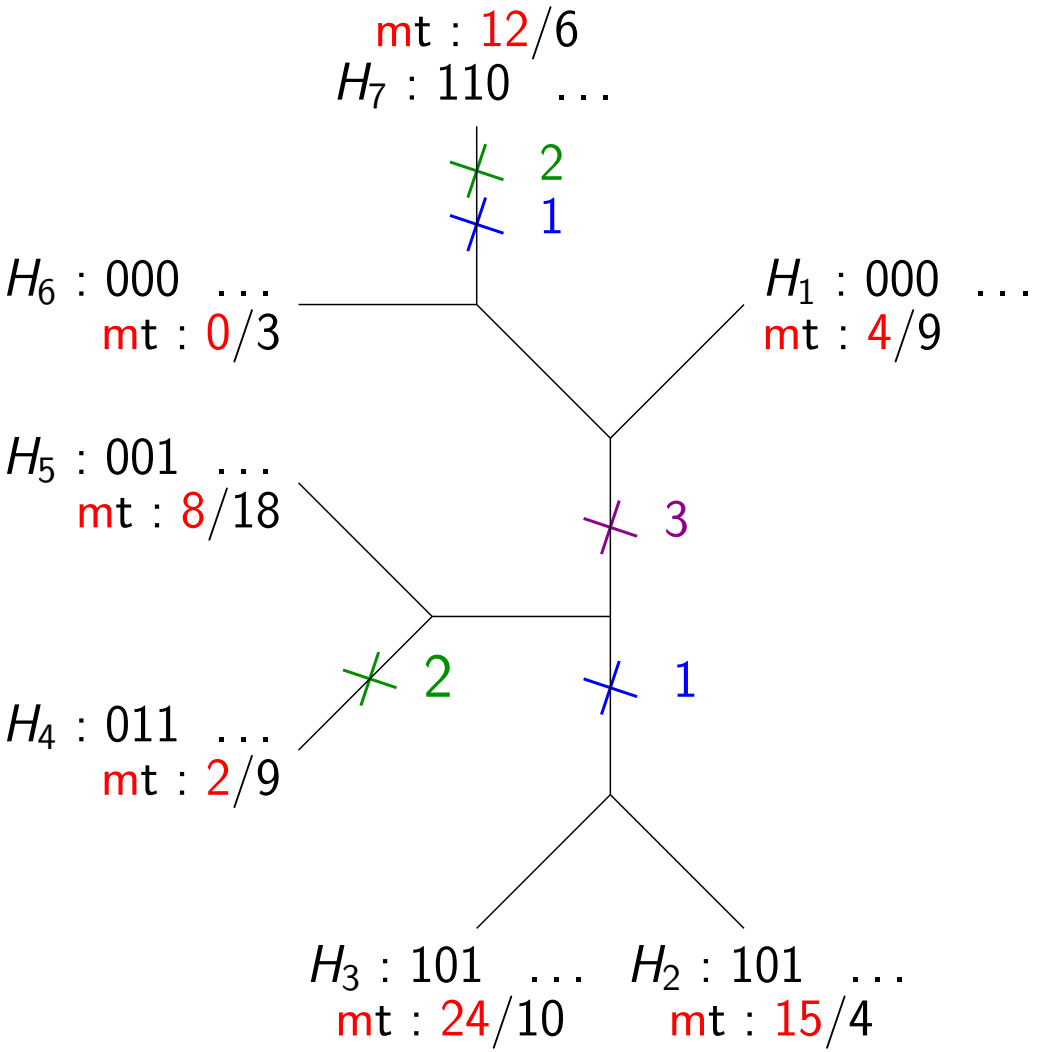
H_2 : 101 ...
mt : 15/4

Données initiales

- 7 haplotypes formés par plusieurs SNPs
- Illustration sur 3 des SNPs
- Haplotypes portés par des malades et des témoins

Méthode

mt : nombre de malades/témoin

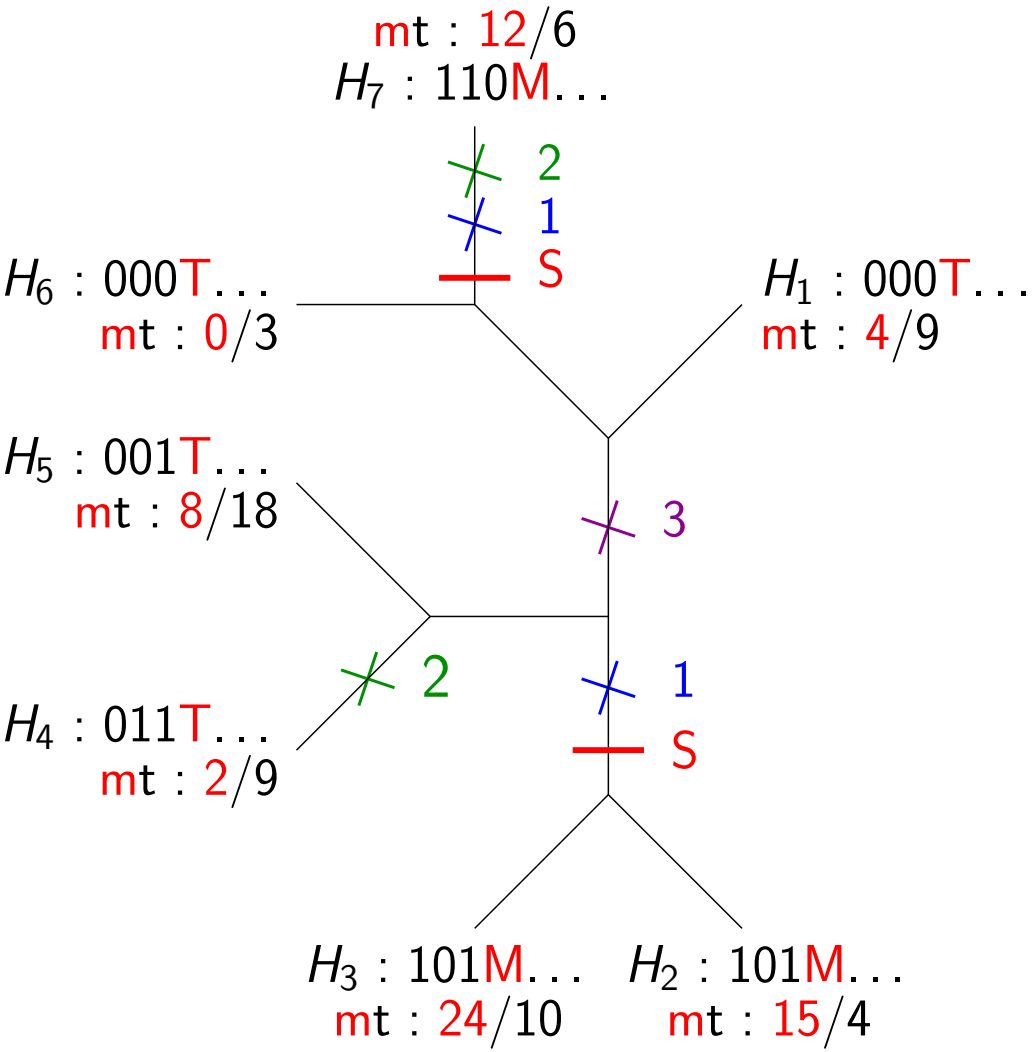


1/ Construction de la phylogénie des haplotypes

- Méthode de parcimonie (ou de maximum de vraisemblance)
- Identification des états de caractères aux nœuds et donc des branches portant des changements d'état

Méthode

mt : nombre de malades/témoin

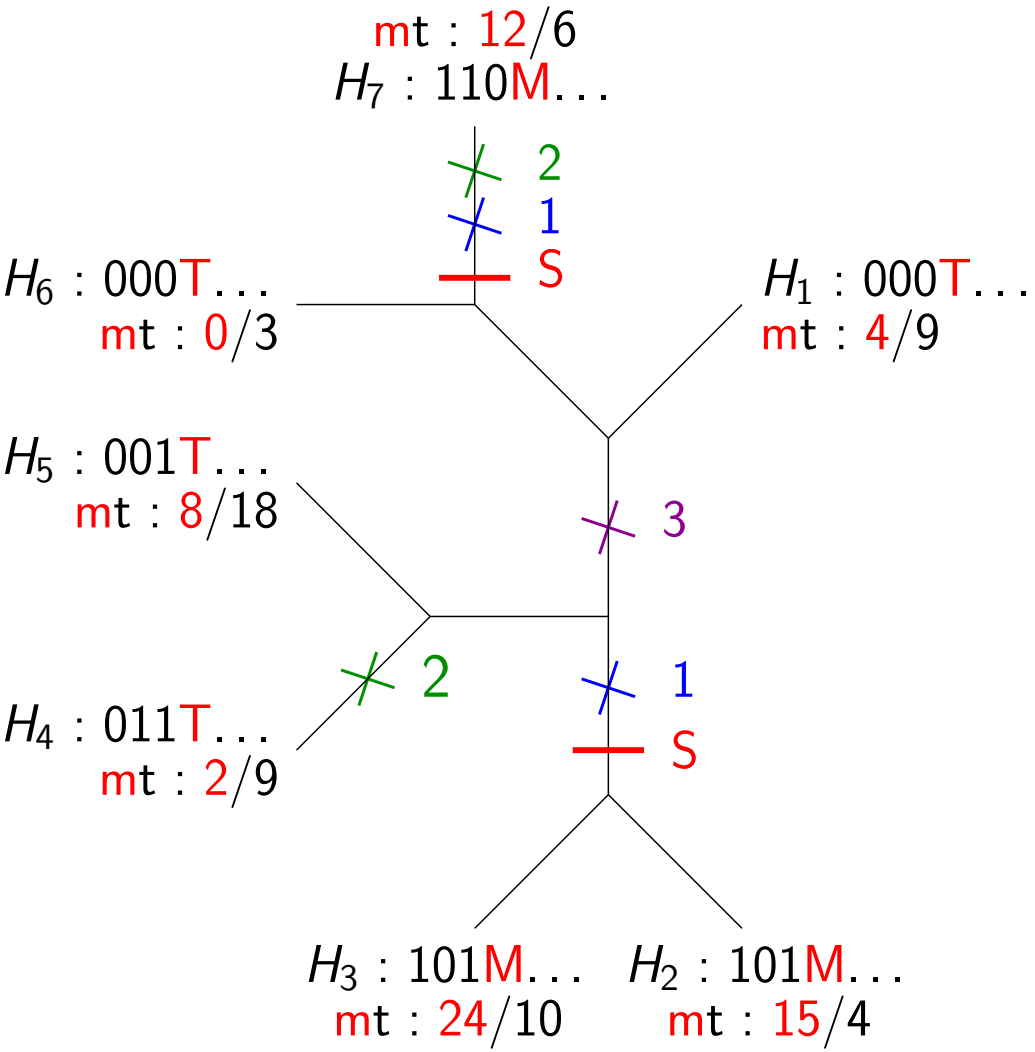


2/ Définition d'un nouveau caractère (S)

- S est défini pour chaque haplotype
- S a 2 états : M et T
- M (resp.T) est attribué aux haplotypes portés par une majorité de malades (resp. témoins)
- Identification des changements d'état du caractère *S* dans l'arbre

Méthode

mt : nombre de malades/témoin



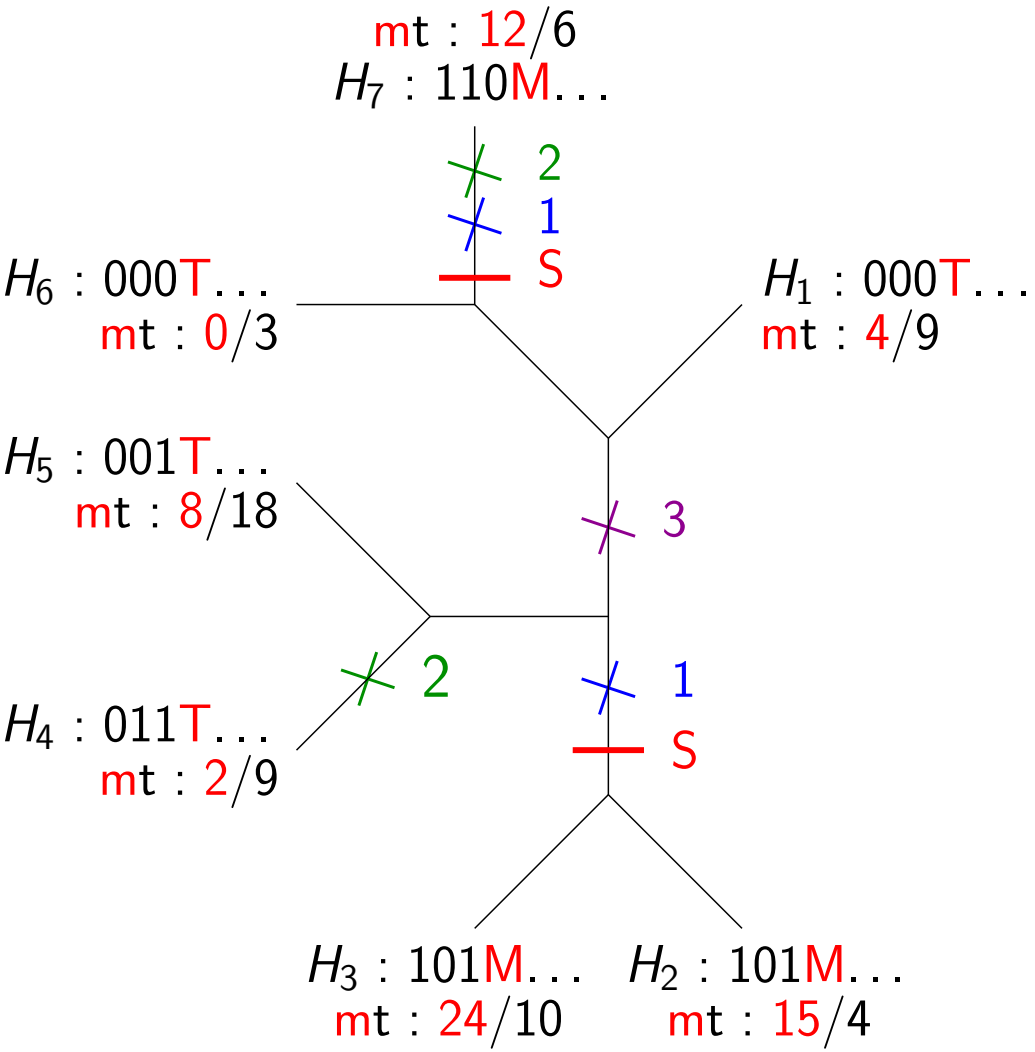
3/ Définition d'un indice de co-évolution : V_i

- O_i : nombre de co-mutations observées pour le SNP i
  $O_1 = 2$; $O_2 = 1$; $O_3 = 0$
- E_i : nombre de co-mutations attendues sous l'hypothèse de distribution aléatoire des mutations sur les 11 branches de l'arbre
  - $E_1 = \frac{2}{11} \times \frac{2}{11} \times 11 = 0.4$
  - $E_2 = 0.4$; $E_3 = 0.2$
- Définition de V_i :

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

Méthode

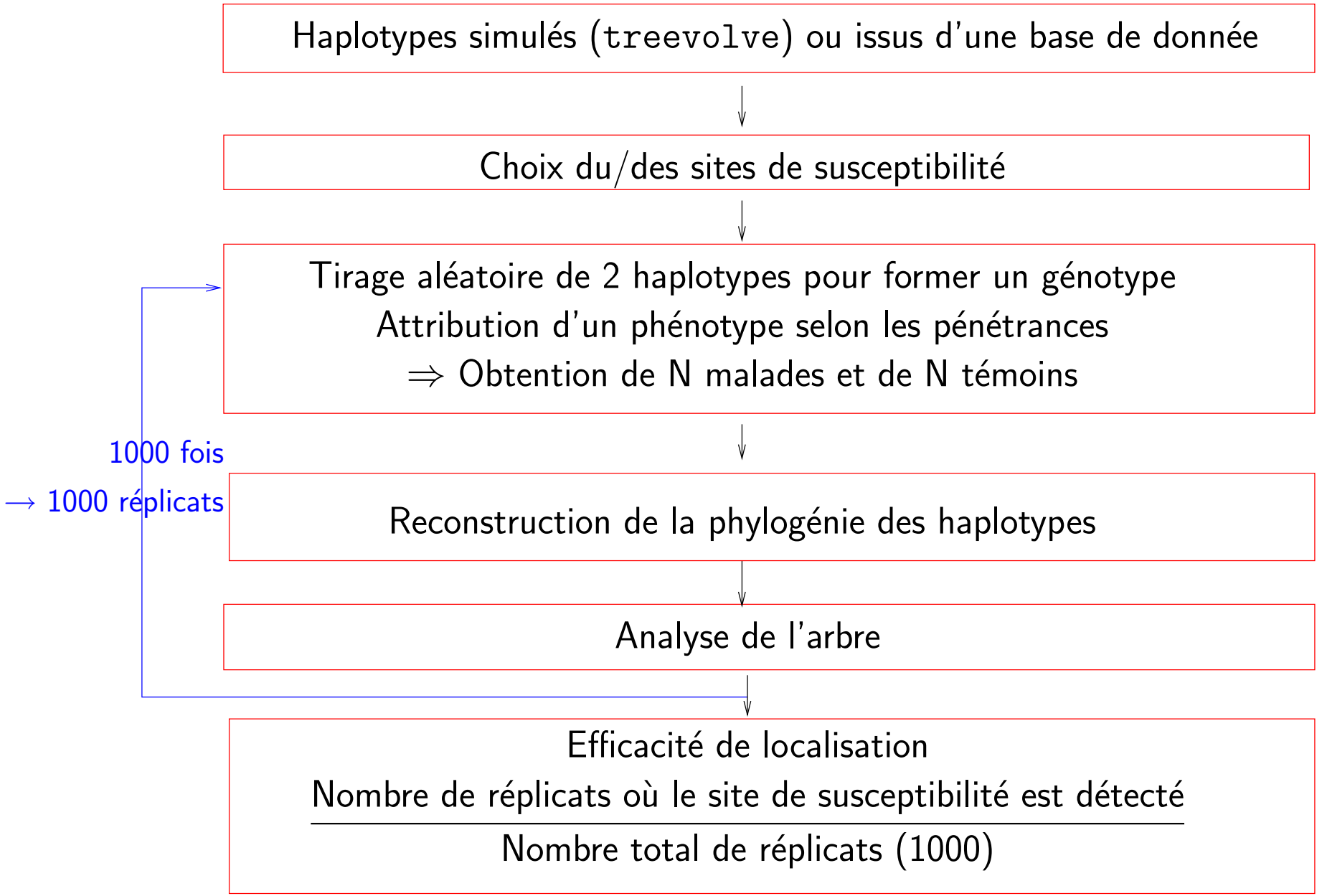
mt : nombre de malades/témoin



4/ Identification des locus à risque

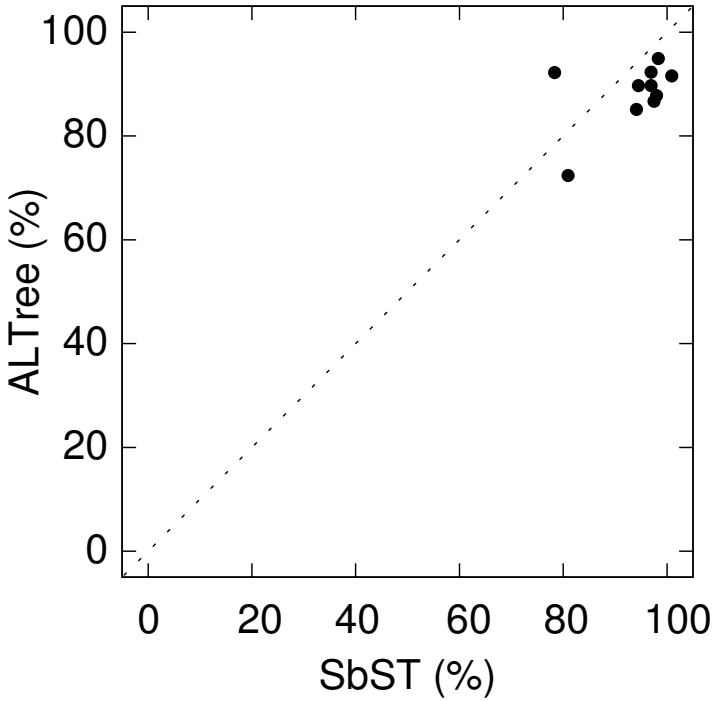
- Calcul de V_i pour tous les SNPs
  - $V_1 = 2.7$ $V_2 = 1.1$ $V_3 = -0.4$
- S'il y a plusieurs arbres équiparcimonieux : additionner les V_i des différents arbres pour chaque site
- Les sites dont les V_i sont les plus élevés sont des sites de susceptibilité potentiels pour la maladie

Processus de simulation



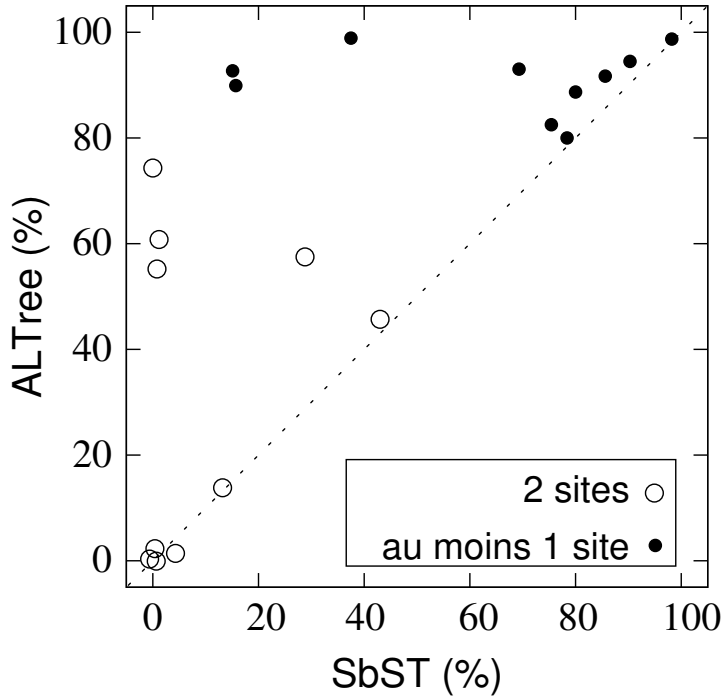
Comparaison avec une méthode locus par locus

1 locus à risque



fréquence : 0.2
risque $\times$ 2 pour les hétérozygotes
risque $\times$ 10 pour les homozygotes

2 locus à risque (allèles A_1 et B_1)

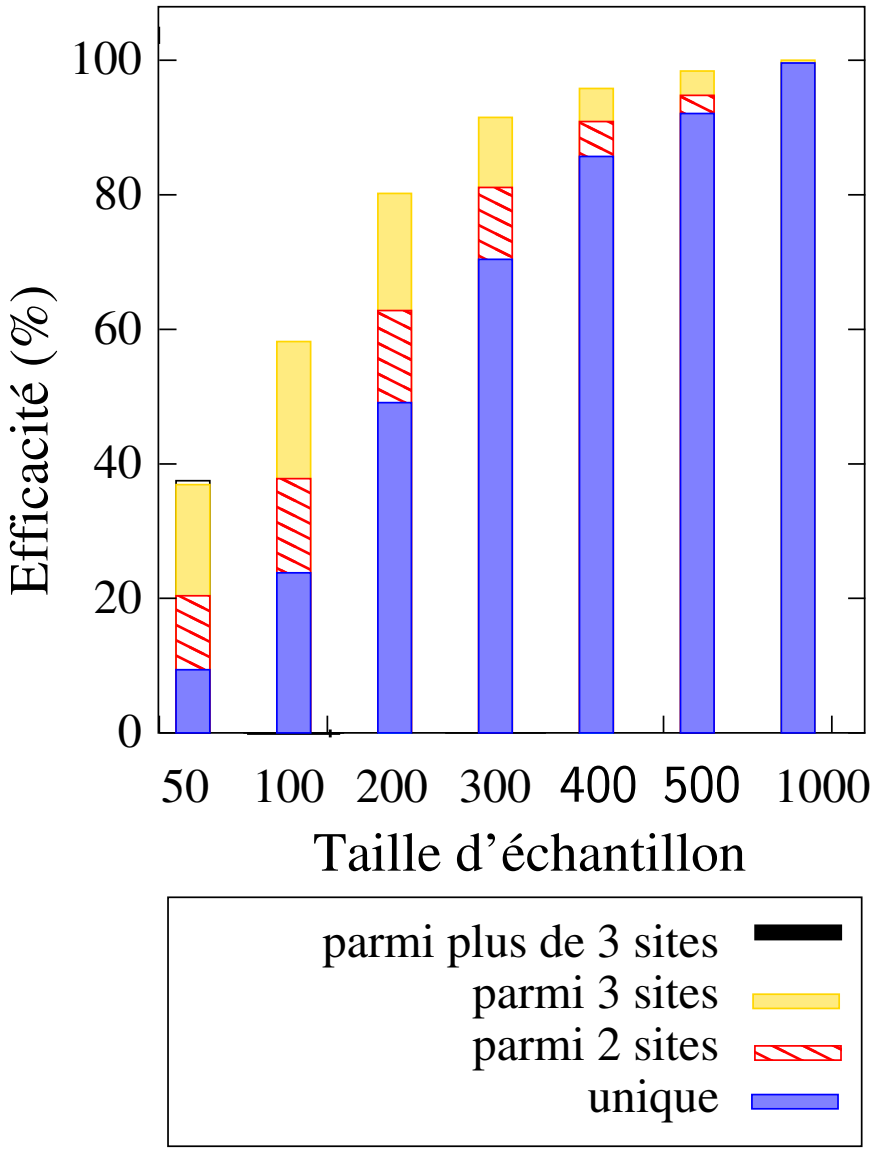


fréquence : $f(A_1B_1) = 0.25$
risque $\times$ 3 pour les porteurs de A_1 et B_1

Conclusion

Méthode surtout intéressante lorsqu'il y a plusieurs locus à risque

Effet de la taille de l'échantillon



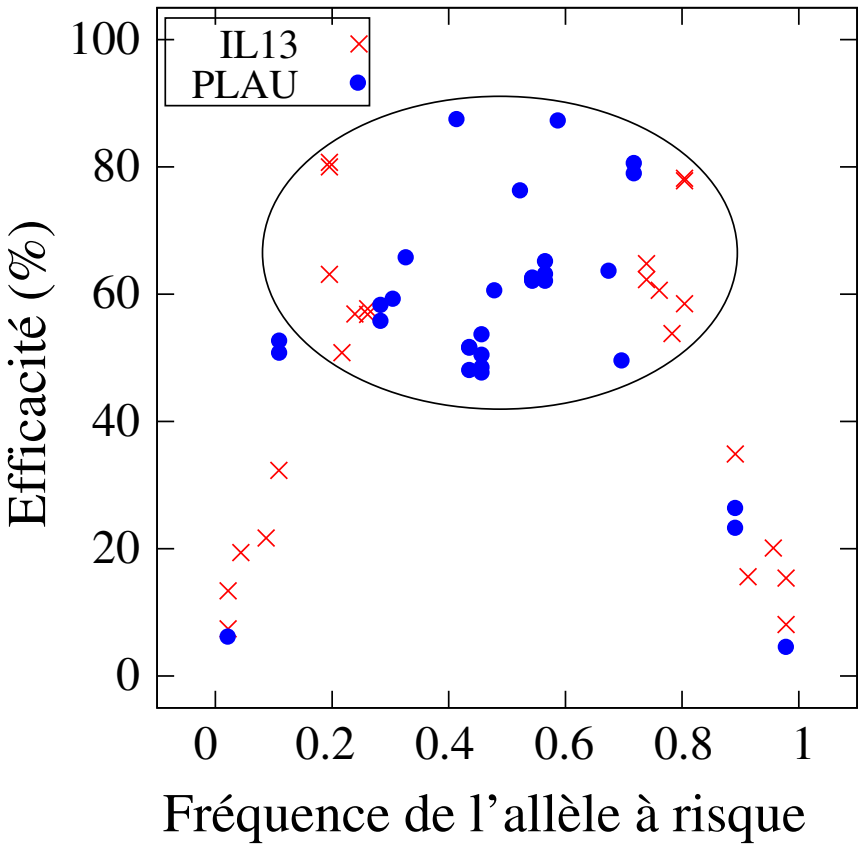
Conclusion

- Comme attendu, augmentation de l'efficacité avec la taille d'échantillon
- Augmentation de la proportion de réplicats dans lesquels le locus à risque est le seul détecté avec la taille de l'échantillon

Modèle génétique simulé :

- freq. de l'allèle à risque : 0.3
- risque des hétérozygotes $\times$ 2
- risque des homozygotes $\times$ 10

Effet de la fréquence f de l'allèle à risque



Taille ech : 200 malades + 200 témoinss
Risques : $\times 2$ pour les hétérozygotes
$\times 10$ pour les homozygotes

Conclusions

Dans nos conditions de simulation :

- efficacité maximale quand $0.2 < f < 0.8$ (maladies complexes)
  - Pas de corrélation entre f et l'efficacité
- Si $f < 0.2$ ou $f > 0.8$, efficacité faible et corrélée avec f
  - Probablement dû aux pénétrances choisies (fort taux de phénocopies)

Application à des données : DRD2 et la schizophrénie

Les données (Dubertret *et al.*, 2004)

- 103 trios (parents + 1 enfant atteint)
  - haplotypes transmis $\longrightarrow$ échantillon de malades
  - haplotypes non transmis $\longrightarrow$ échantillon de contrôles
- 7 SNPs + 1 microsatellite génotypés dans DRD2 et la kinase X
- Haplotypes reconstruits avec `FAMHAP` v1.5 $\longrightarrow$ 59 haplotypes $\neq$
- 1000 arbres équiparcimonieux reconstruits avec `PAUP`

Résultats

- Les V_i sont calculés pour les 7 SNPs mais pas pour le microsatellite
- Le SNP dont le V_i est le plus élevé est le **SNP 3 (TaqI A1/A2)**, A2 étant l'allèle à risque
- Ce SNP a déjà été identifié comme étant à risque pour la schizophrénie (Dubertret *et al.*, 2001, 2004)

Conclusion

Développement d'une nouvelle méthode

- Méthode de localisation des sites de susceptibilité
- Implémentation dans le logiciel `ALTree` (`http://claire.bardel.free.fr/software.html`)
- Méthode facilement applicable à des données quantitatives et à la recherche de QTN (quantitative trait nucléotide)

Étude de son efficacité

- Plus efficace que les méthodes locus par locus quand plusieurs sites sont à risque
- Efficacité dépendante de la taille de l'échantillon
- Efficacité maximale quand f est comprise entre 0.2 et 0.8
- Autres études : l'efficacité dépend aussi des pénétrances et de la position du locus à risque dans l'arbre

Perspectives

Le problème de la recombinaison

- Méthode valable dans des blocs haplotypiques
  $\longrightarrow$ Tester l'influence de la recombinaison en simulation

Augmenter les fonctionnalités de la méthode

- Prendre en compte l'incertitude liée à la reconstruction haplotypique
- En ML : prendre en compte les probabilités des états de caractère aux nœuds dans le calcul du V_i

Appliquer la méthode à d'autres jeux de données

- Données qualitatives sur la schizophrénie : nouveaux marqueurs en cours de génotypage dans la kinase X
- Données quantitatives sur le taux de TAFI (Thrombin-Activatable Fibrinolysis Inhibitor)

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 :

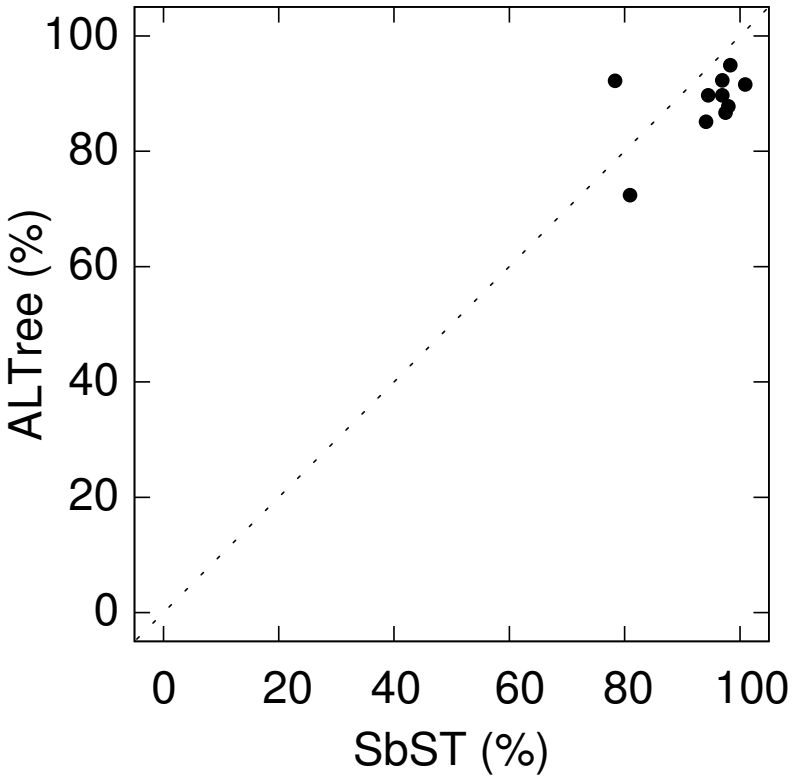
$$E_i^{0\to1} = \frac{(m_i^{0\to1} \times s^{T\to M}) + (m_i^{1\to0} \times s^{M\to T})}{b}$$

$m_i^{0\to1}$: 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

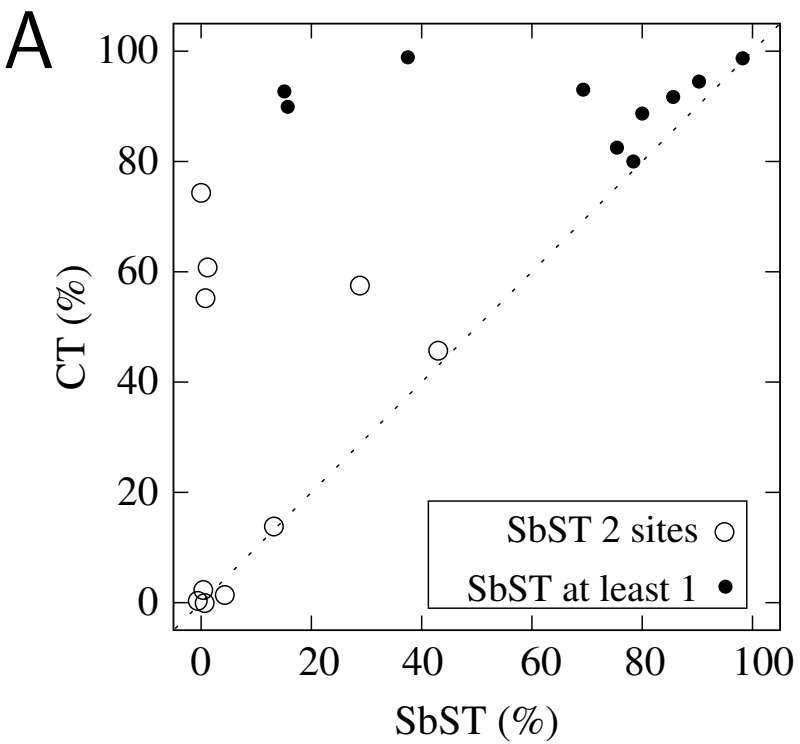
$$\Rightarrow \quad \begin{cases} V_i^{0\to1} = 0 & \text{if } E_i^{0\to1} = 0 \\ V_i^{0\to1} = \dfrac{R_i^{0\to1} - E_i^{0\to1}}{\sqrt{E_i^{0\to1}}} & \text{if } E_i^{0\to1} \neq 0 \end{cases}$$

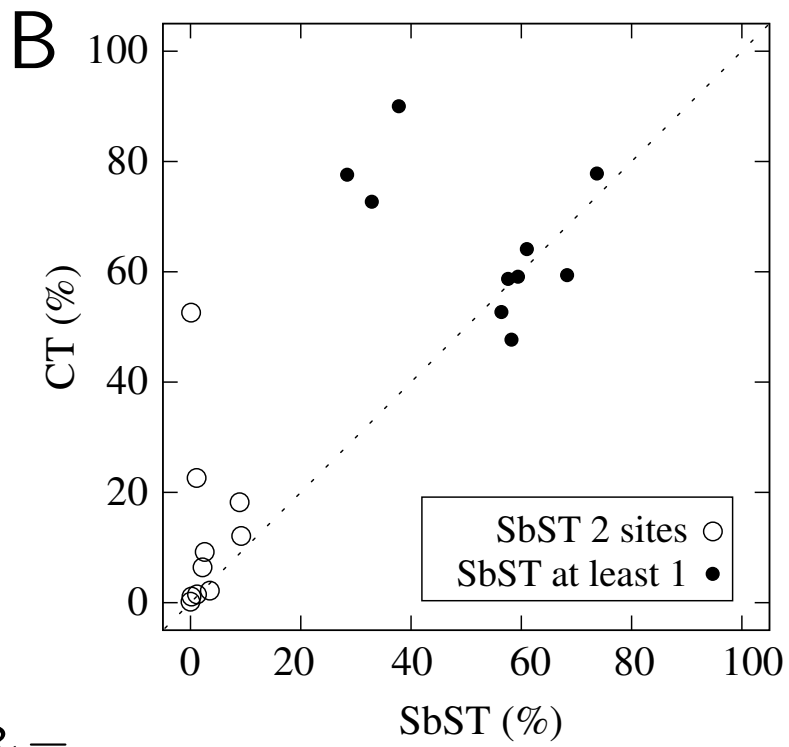
Comparison with a locus by locus method



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 $p = 0.3$