Mise en évidence de facteurs génétiques de risque en utilisant des phylogénies d'haplotypes

Claire Bardel

Thèse réalisée à l'unité INSERM U535, Villejuif, France
sous la direction de Pierre Darlu et Emmanuelle Génin

12 décembre 2005

Introduction générale

Contexte des travaux

- Recherche de facteurs génétiques de risque
- Maladies multifactorielles
  - dues à des facteurs multiples (génétiques, environnementaux)
  - exemples : maladies cardio-vasculaires, neurologiques,...
- Développement des méthodes de biologie moléculaire

  $\longrightarrow$ Augmentation du nombre de données disponibles
- Intérêt de développer de nouvelles méthodes et stratégies pour utiliser toute cette information

But de la thèse

- Développement et étude des propriétés d'une nouvelle méthode de recherche de facteurs génétiques de risque
  - Méthode basée sur l'utilisation de l'histoire évolutive des séquences étudiées

Plan de l'exposé

Les outils utilisés

Utilisation de marqueurs génétiques : les SNPs

- Changement ponctuel d'un seul nucléotide
- Très nombreux ($>$ 10 millions sur le génome humain), faciles à génotyper et situés sur tout le génome
- bi-alléliques : codage 0/1

Utilisation conjointe de plusieurs marqueurs

- Définition d'**haplotypes** : ensemble des allèles des différents marqueurs situés sur un même chromosome
- Reconstruction des haplotypes par des méthodes d'inférence statistique

Les deux grands types de méthodes

Deux stratégies d'analyse

- Recherche globale sur le génome
- Gène candidat

Deux types de méthodes

- Les analyses de liaison
  - Analyse de données familiales
  - Recherche de marqueurs dont la transmission dans les familles n'est pas indépendante de la maladie
- Les analyses d'association
  - Analyse de données **en population**
  - Comparaison d'échantillons de **malades** et de **témoins**

Test sur des données en population

Le test site par site (SbST)

- Soit un locus bi-allélique (A_1/A_2)
- Test de l'homogénéité de distribution des allèles chez les malades et les témoins (χ^2 à 1 ddl)

	A_1	A_2	Total
Malades	m_1	m_2	M
Témoins	t_1	t_2	T
Total	N_1	N_2	N

- Correction pour les tests multiples
- Puissance maximale quand le variant est un des locus étudiés ou est en fort DL avec un des locus

Extension à d'autres types de données

- données génotypiques ou haplotypiques

Problèmes des tests haplotypiques

Problèmes liés à l'augmentation du nombre de marqueurs

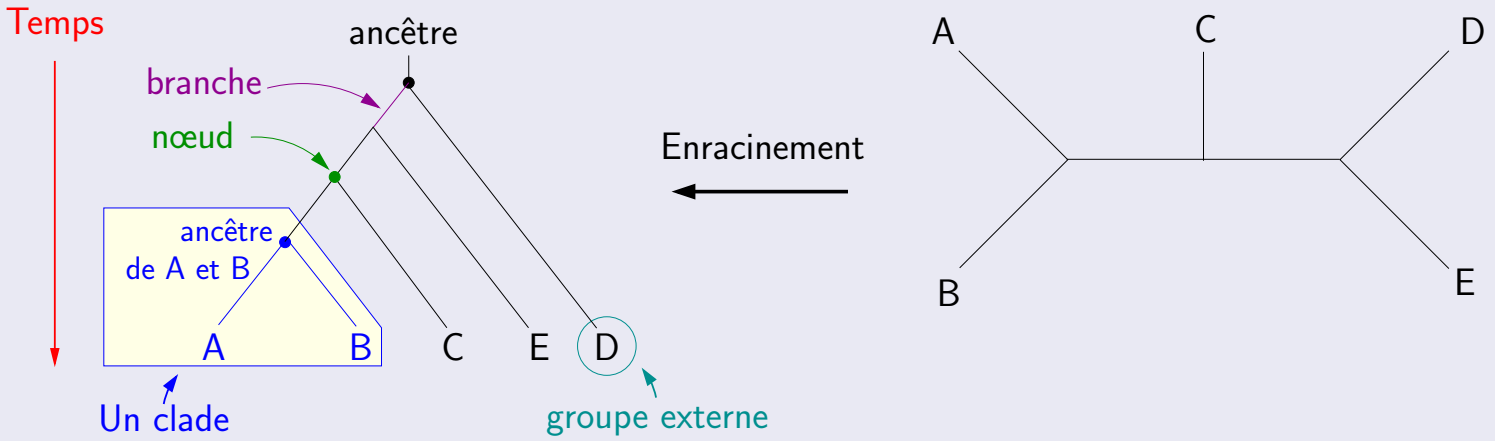
- Augmentation du nombre d'haplotypes possibles
  $\longrightarrow$ baisse de puissance du test
- Diminution de l'effectif pour chaque haplotype
  $\longrightarrow$ problème de petits échantillons (surtout pour les χ^2)

Une solution : le regroupement de catégories

- Regroupement des haplotypes rares en une catégorie
- Regroupement des haplotypes rares avec les haplotypes qui leur ressemblent le plus
- Regroupement selon un arbre de classification
- Regroupement selon l'histoire évolutive des haplotypes

Les phylogénies

Représentation de l'histoire évolutive : les phylogénies



Phylogénie ou arbre enraciné

Arbre non enraciné

L'enracinement

- Utilisation d'un groupe externe
- Choix d'un ancêtre
  - Haplotype le plus fréquent
  - Haplotype consensus : allèles les plus fréquents à chaque locus

Utilisation des phylogénies en génétique épidémiologique

101001... haplo 1 (10 témoins **1 malade**)

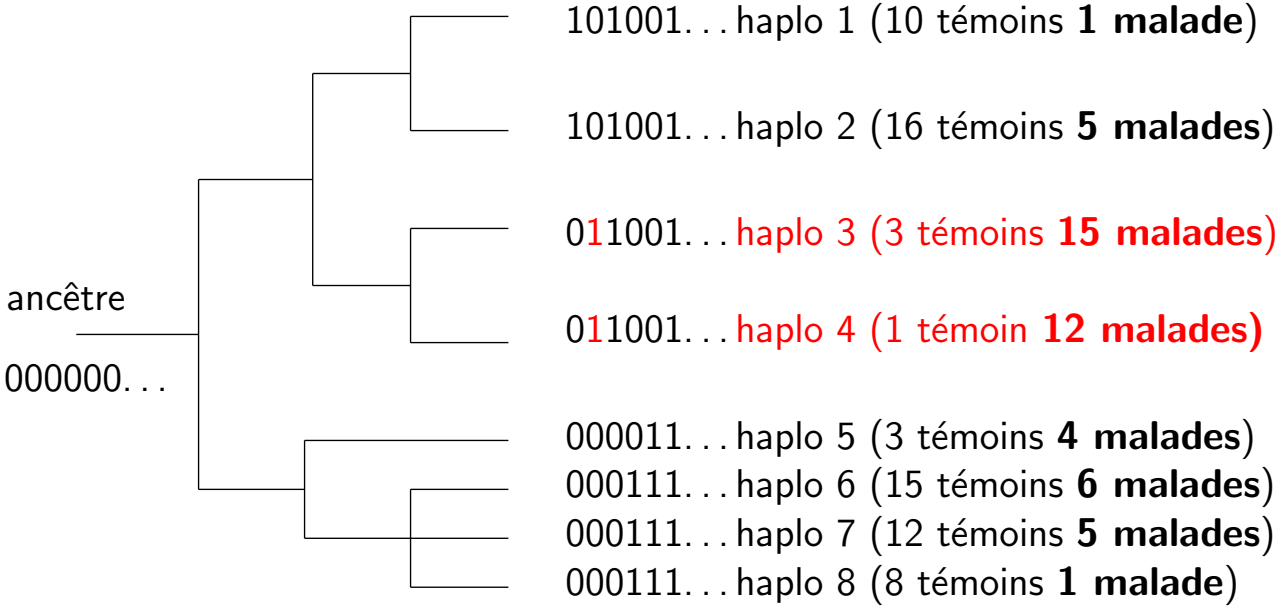
101001... haplo 2 (16 témoins **5 malades**)

011001... haplo 3 (3 témoins **15 malades**)

011001... haplo 4 (1 témoin **12 malades)**

000011... haplo 5 (3 témoins **4 malades**)
000111... haplo 6 (15 témoins **6 malades**)
000111... haplo 7 (12 témoins **5 malades**)
000111... haplo 8 (8 témoins **1 malade**)

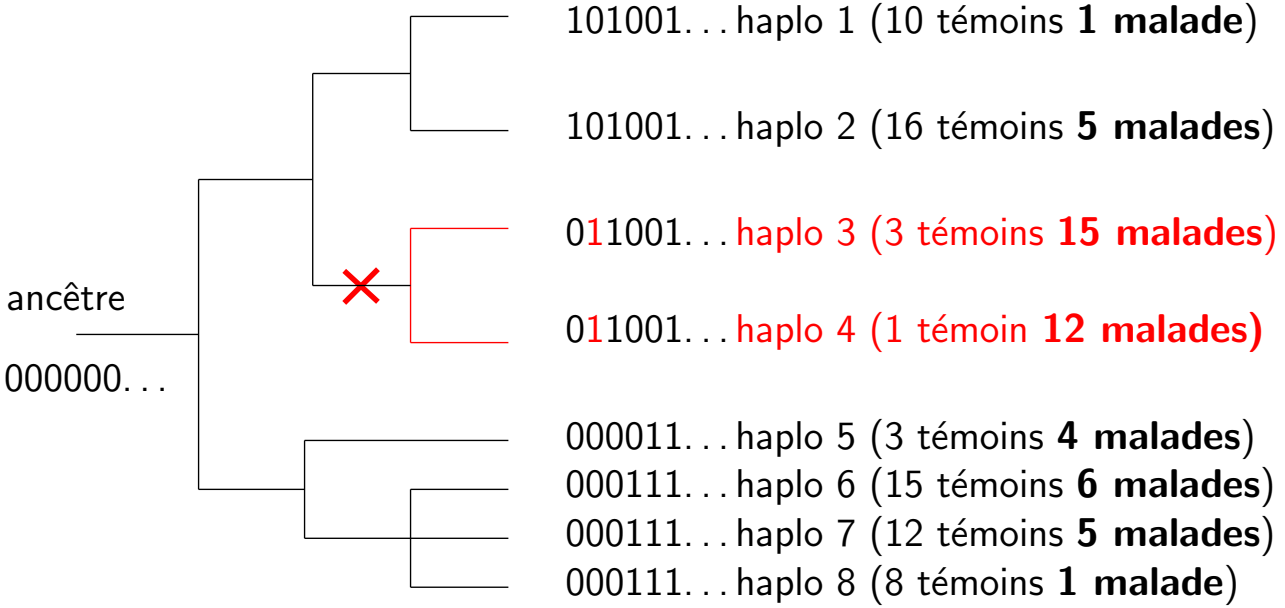
Utilisation des phylogénies en génétique épidémiologique



Comment utiliser des phylogénies en génétique épidémiologique ?

- Regroupement des haplotypes $\longrightarrow$ Test d'association

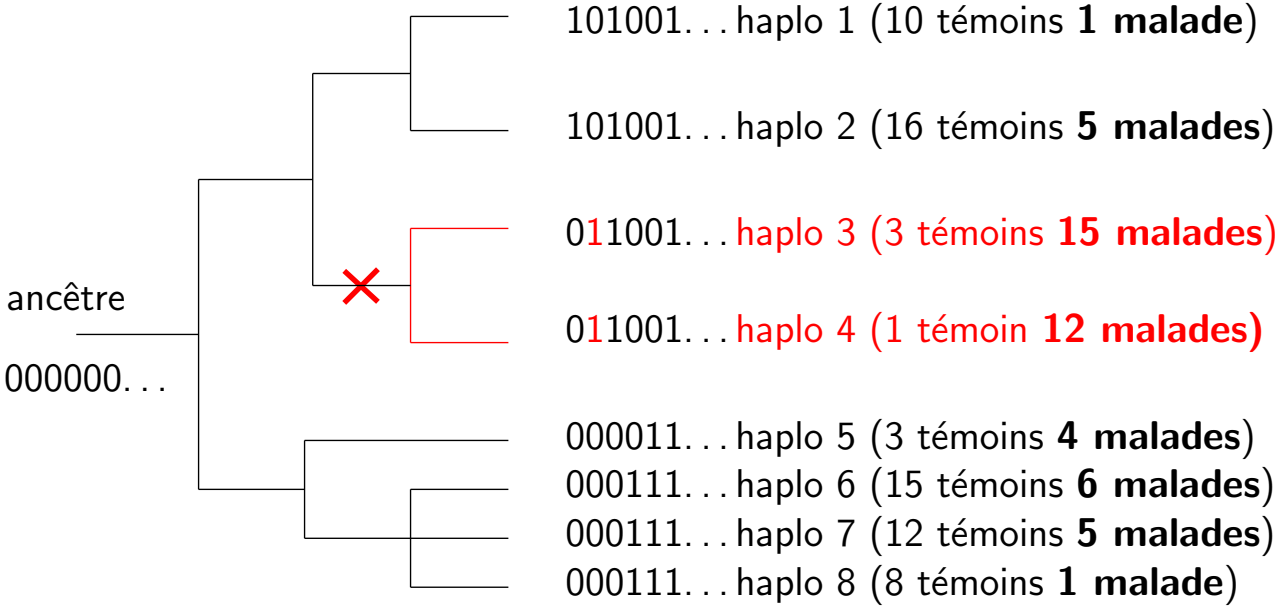
Utilisation des phylogénies en génétique épidémiologique



Comment utiliser des phylogénies en génétique épidémiologique ?

- Regroupement des haplotypes $\longrightarrow$ Test d'association
- Recherche de clades contenant significativement plus de malades $\longrightarrow$ Identification du/des sites de susceptibilité

Utilisation des phylogénies en génétique épidémiologique



Comment utiliser des phylogénies en génétique épidémiologique ?

- Regroupement des haplotypes → Test d'association
- Recherche de clades contenant significativement plus de malades → Identification du/des sites de susceptibilité
- Plusieurs méthodes développées, la plupart durant ma thèse

Plan de l'exposé

Généralités

Les données

- Données haplotypiques
- Courtes séquences de marqueurs bi-alléliques

Reconstruction de l'arbre

- Par parcimonie (PAUP, PHYLIP)
- Par maximum de vraisemblance (PHYML)

Exemple

Exemple : 40 témoins, 40 malades, 5 haplotypes formés de 5 SNPs + une séquence ancestrale

H001	10110	2 t	**10 m**
H002	10100	1 t	**5 m**
H003	11100	5 t	**12 m**
H004	01101	22 t	**13 m**
H005	01110	10 t	**0 m**

Généralités

Les données

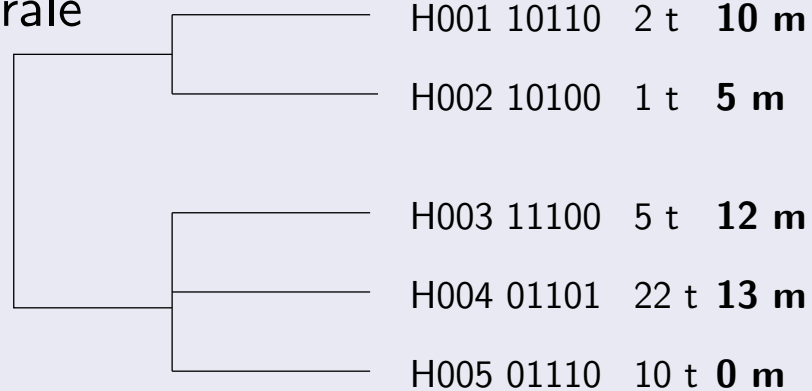
- Données haplotypiques
- Courtes séquences de marqueurs bi-alléliques

Reconstruction de l'arbre

- Par parcimonie (PAUP, PHYLIP)
- Par maximum de vraisemblance (PHYML)

Exemple

Exemple : 40 témoins, 40 malades, 5 haplotypes formés de 5 SNPs + une séquence ancestrale

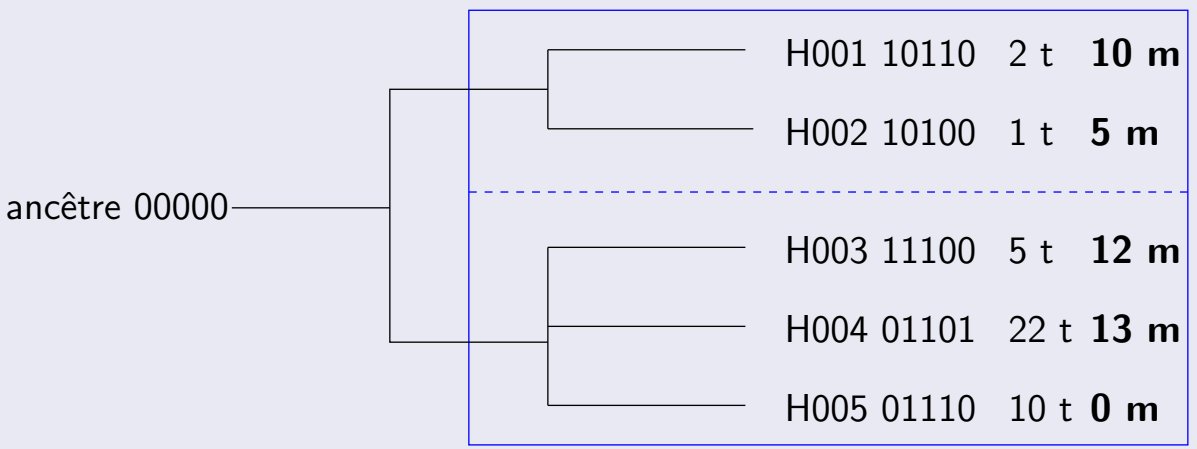


Le test d'association (1)

Enracinement de l'arbre

- Parcours de l'arbre depuis la racine → Définition non ambiguë des groupes

Analyse emboîtée sur l'arbre



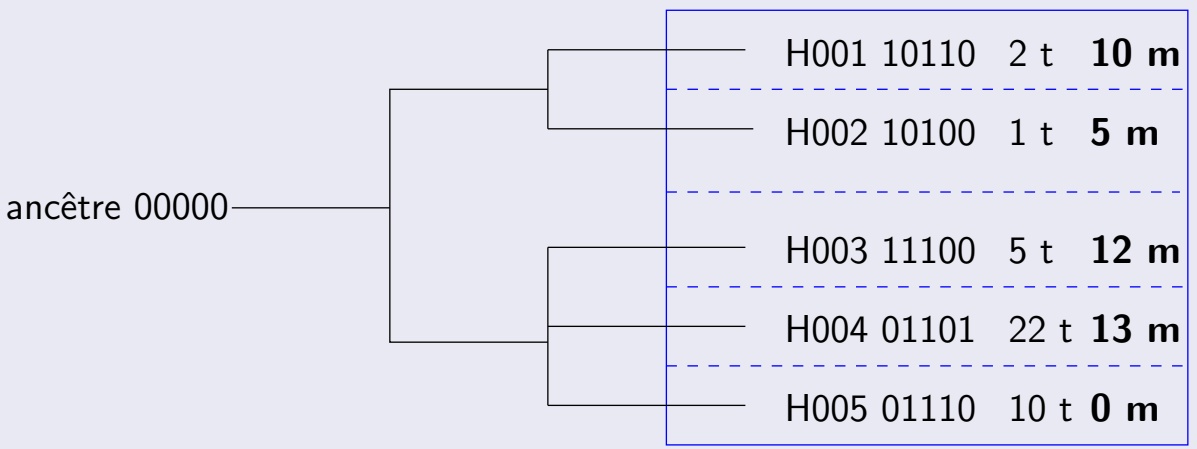
- Un test à 1 ddl : χ^2=10

Le test d'association (1)

Enracinement de l'arbre

- Parcours de l'arbre depuis la racine → Définition non ambiguë des groupes

Analyse emboîtée sur l'arbre



- Un test à 1 ddl : χ^2=10
- Un test à 4 ddl : χ^2=23

Correction pour les tests multiples

		Test 1	Test 2
Statistique observée		10	23
Simulés sous H_0	Réplicat 1	12,3	23,9
	Réplicat 2	8,9	21,8
	Réplicat 3	11,8	20,3
	Réplicat 4	9,4	19,7
	Réplicat 5	10,5	22,4

Étape 1 :

Calcul des valeurs de la statistique pour le jeu de donnée et pour les réplicats simulés

Correction pour les tests multiples

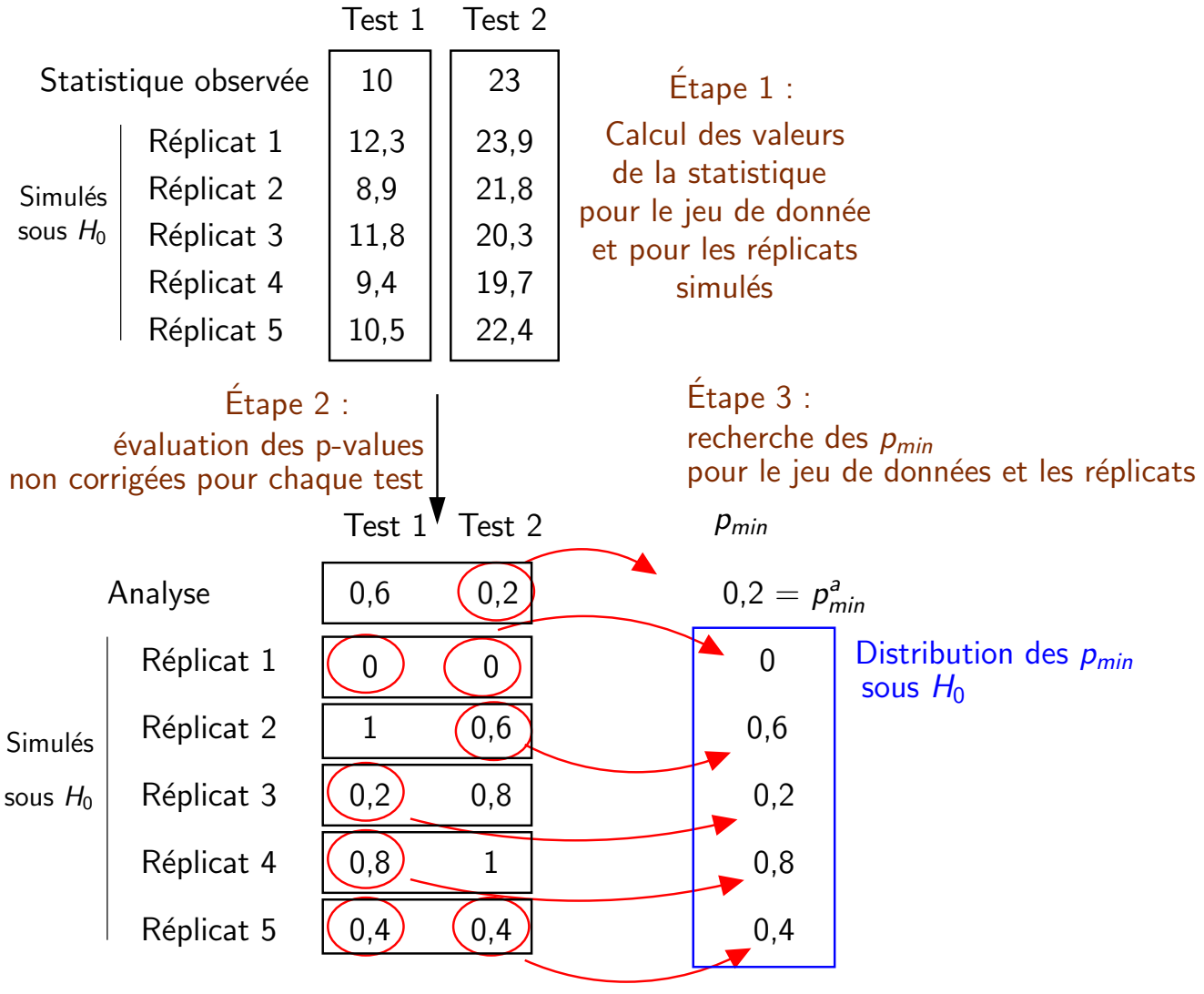
		Test 1	Test 2
Statistique observée		10	23
Simulés sous H_0	Réplicat 1	12,3	23,9
	Réplicat 2	8,9	21,8
	Réplicat 3	11,8	20,3
	Réplicat 4	9,4	19,7
	Réplicat 5	10,5	22,4

Étape 1 :
Calcul des valeurs de la statistique pour le jeu de donnée et pour les réplicats simulés

Étape 2 :
évaluation des p-values non corrigées pour chaque test

		Test 1	Test 2
Analyse		0,6	0,2
Simulés sous H_0	Réplicat 1	0	0
	Réplicat 2	1	0,6
	Réplicat 3	0,2	0,8
	Réplicat 4	0,8	1
	Réplicat 5	0,4	0,4

Correction pour les tests multiples



Correction pour les tests multiples

		Test 1	Test 2
Statistique observée		10	23
Simulés sous H_0	Réplicat 1	12,3	23,9
	Réplicat 2	8,9	21,8
	Réplicat 3	11,8	20,3
	Réplicat 4	9,4	19,7
	Réplicat 5	10,5	22,4

Étape 1 :
Calcul des valeurs de la statistique pour le jeu de donnée et pour les réplicats simulés

Étape 2 :
évaluation des p-values non corrigées pour chaque test

		Test 1	Test 2
Analyse		0,6	0,2
Simulés sous H_0	Réplicat 1	0	0
	Réplicat 2	1	0,6
	Réplicat 3	0,2	0,8
	Réplicat 4	0,8	1
	Réplicat 5	0,4	0,4

Étape 3 :
recherche des p_{min} pour le jeu de données et les réplicats

p_{min}

$0,2 = p^a_{min}$

Distribution des p_{min} sous H_0
0
0,6
0,2
0,8
0,4

Étape 4 :
évaluer la p-value min corrigée
$p = \frac{2}{5} = 0,4$

Identification des locus de susceptibilité : principe

Principe

Pour chaque site, définition d'un indice de co-évolution entre :

- le site
- un nouveau caractère correspondant au statut maladie des haplotypes

Plus l'évolution est corrélée, plus la probabilité que le site soit impliqué dans le déterminisme de la maladie est élevée

Identification des locus de susceptibilité

$H_7 : 110 \quad \ldots$

$H_6 : 000 \quad \ldots$

$H_1 : 000 \quad \ldots$

$H_5 : 001 \quad \ldots$

$H_4 : 011 \quad \ldots$

$H_3 : 101 \quad \ldots$ $H_2 : 101 \quad \ldots$

Données initiales

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

Identification des locus de susceptibilité

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

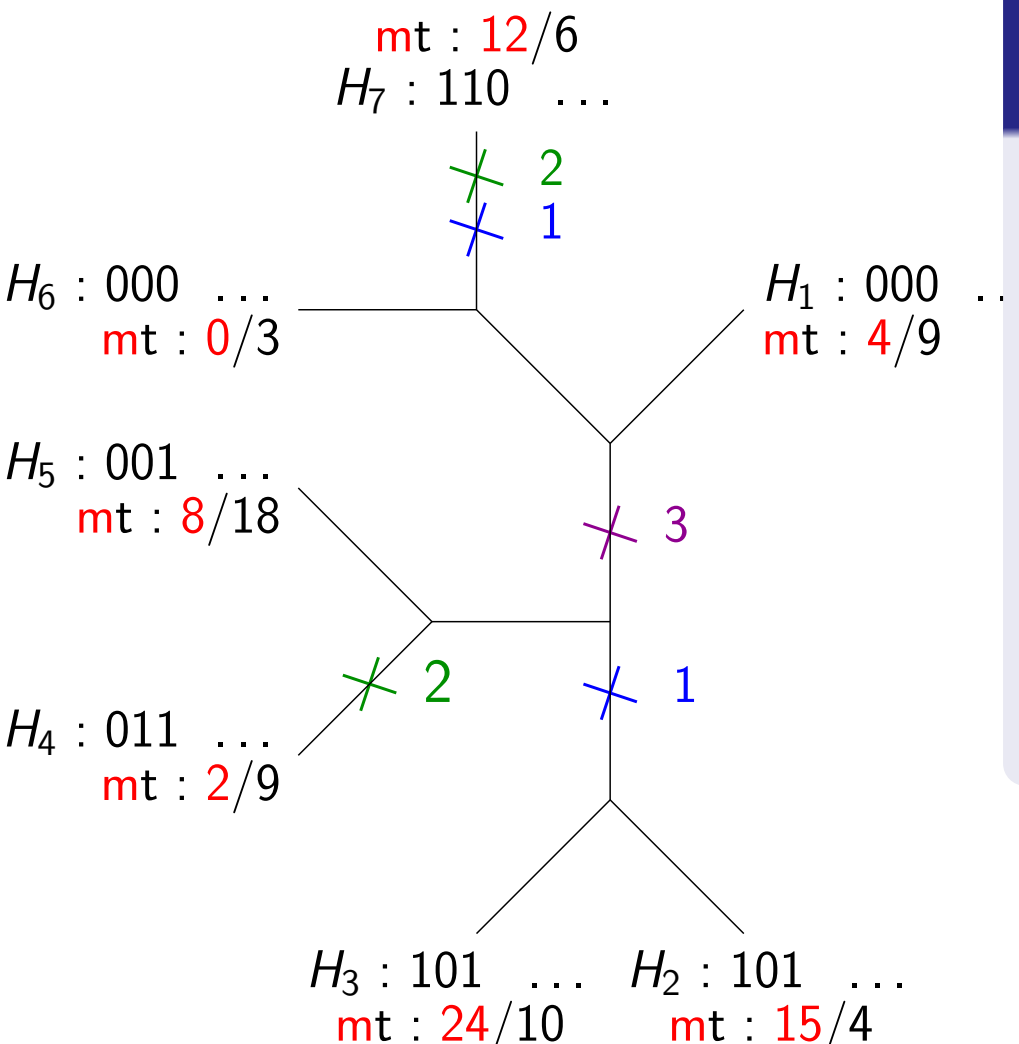
H_2 : 101 ...
mt : 15/4

Données initiales

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

Identification des locus de susceptibilité

mt : nombre de malades/témoin

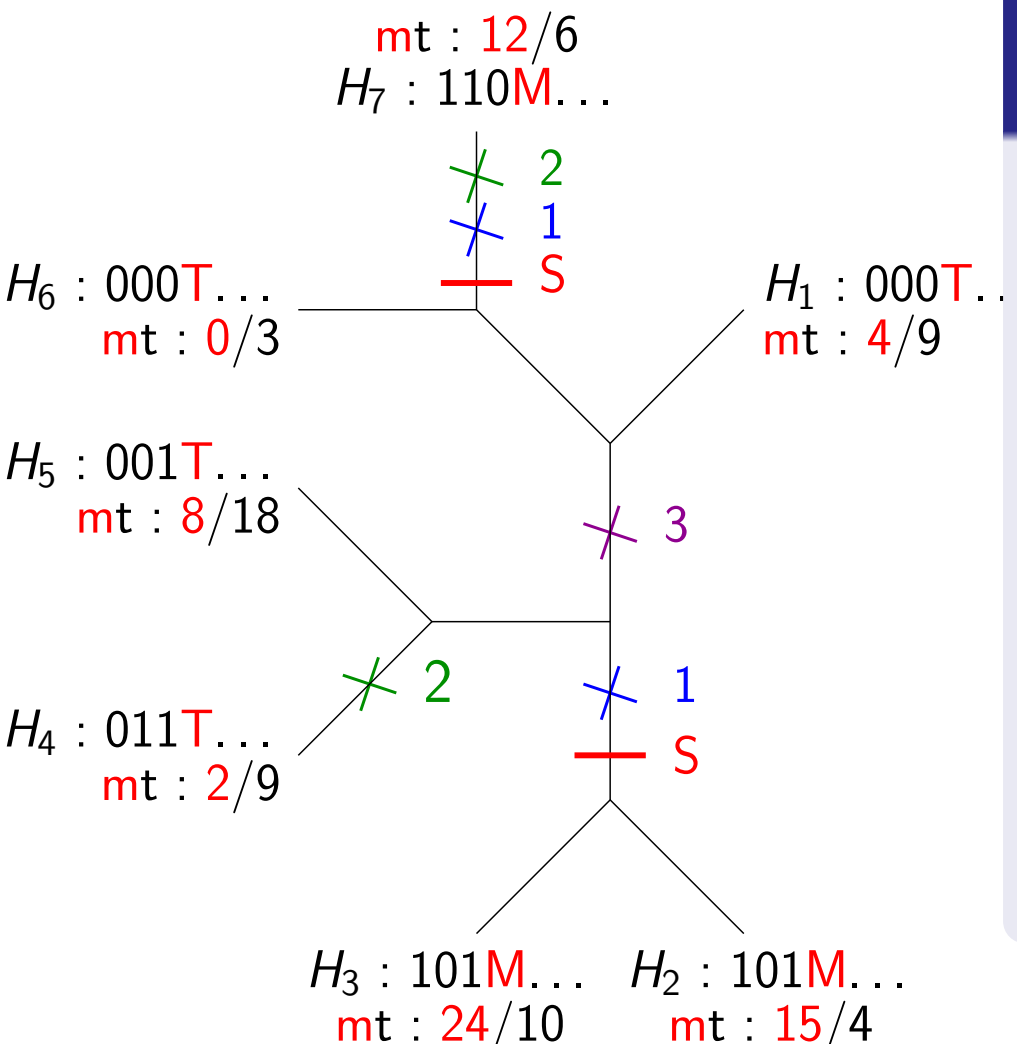


1/ Construction de la phylogénie des haplotypes

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

Identification des locus de susceptibilité

mt : nombre de malades/témoin

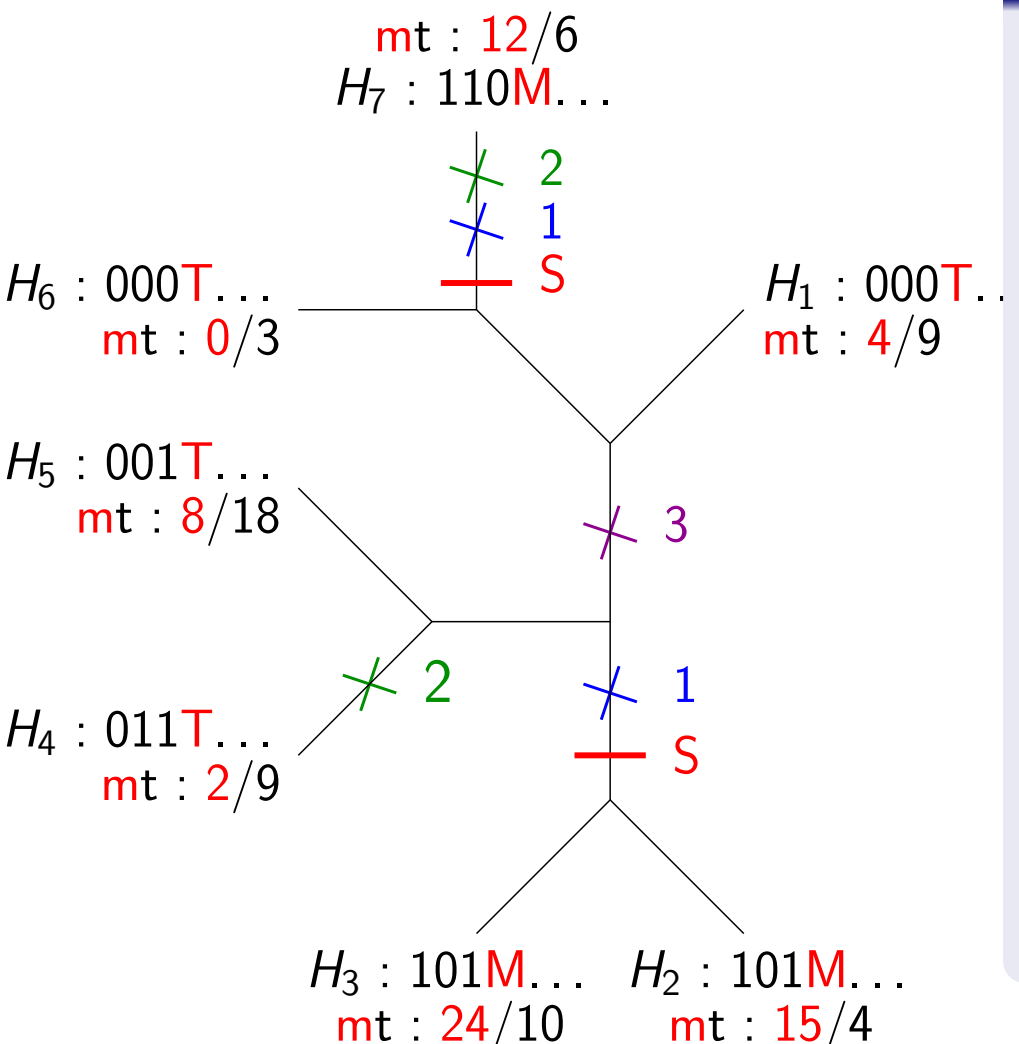


2/ Définition d'un nouveau caractère (S)

- 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

Identification des locus de susceptibilité

mt : nombre de malades/témoin



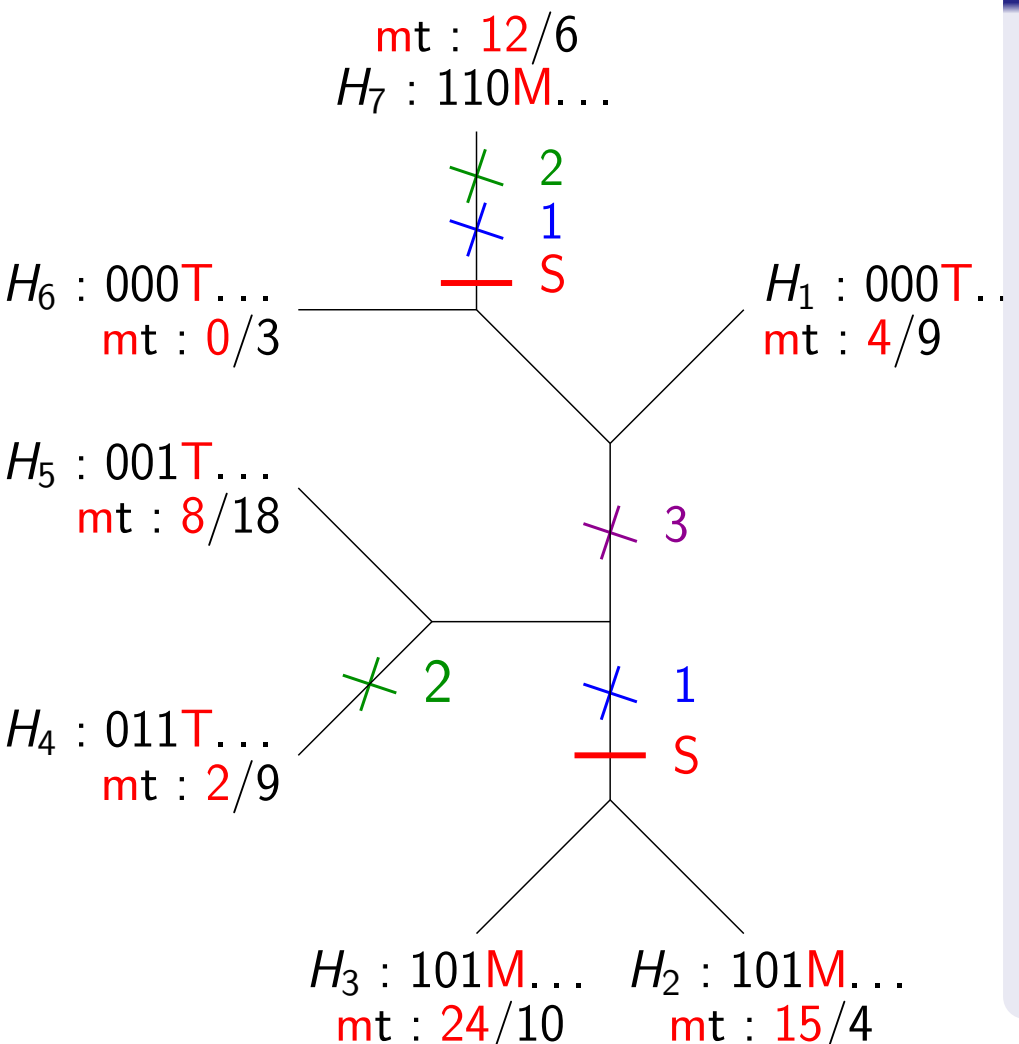
3/ Définition d'un indice de co-évolution : V_i

- O_i : nb co-mutations observées pour le SNP i
  $O_1 = 2$; $O_2 = 1$; $O_3 = 0$
- E_i : nb 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 :

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

Identification des locus de susceptibilité

mt : nombre de malades/témoin

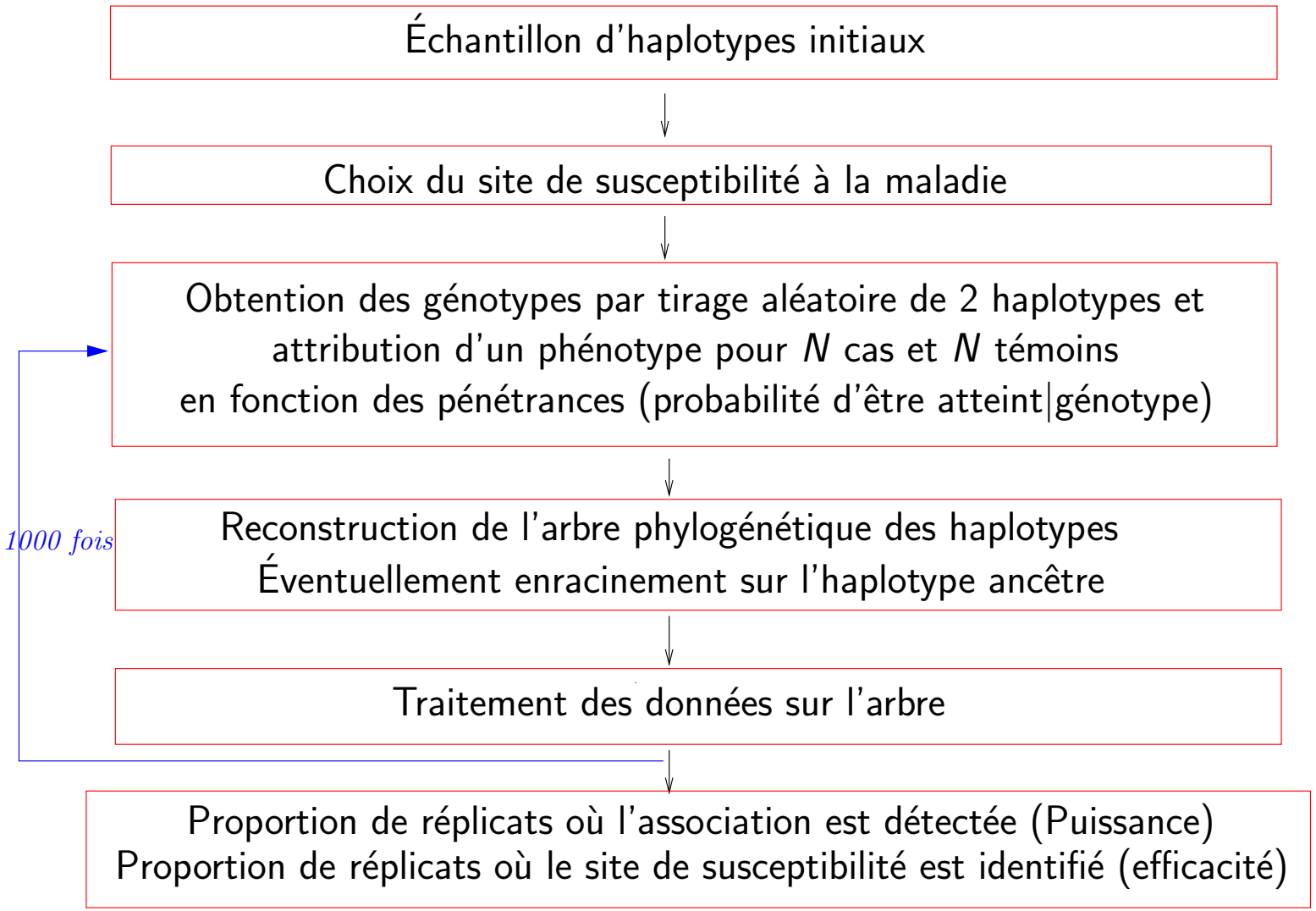


4/ Identification des locus à risque

- 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

Étude de l'efficacité de la méthode par simulations

(Bardel et al, BMC Genet. 2005 ;6(1) :24)



Constitution des échantillons d'haplotypes initiaux

Avec TREEVOLVE (Grassly et Rambaut, 2000)

- Simulation d'un arbre-guide (méthode de coalescence)
- Simulation d'une séquence ancestrale
- Simulation de l'évolution de la séquence ancestrale sur l'arbre-guide
- Choix d'un ou deux sites de susceptibilité en fonction de leurs fréquences alléliques

À partir d'haplotypes issus de banques de données

- Déséquilibres de liaison réalistes entre les marqueurs
- Gènes IL13 et PLAU de la base de données SeattleSNPs :
  - courtes séquences (<10 kb)
  - 1 marqueur tous les 0,5 kb
  - environ 10 haplotypes

Puissance de détection de l'association

Paramètres des simulations (10 jeux de données)

- 4 conditions de simulations :
  - 1 ou 2 sites de susceptibilité
  - Site de susceptibilité inclus ou non parmi les marqueurs étudiés
  - Variation des pénétrances
- Comparaison ALTree, test haplotypique, test SbS

Résultats (1 site de susceptibilité simulé)

	ALTree	HT	SbST
1 site inclus parmi les marqueurs étudiés			
moyenne	92.9 *(3.5)*	91.2 *(4.8)*	99.35 *(4.4)*
écart-type	5.63	0.41	0.13
1 site non inclus parmi les marqueurs étudiés			
moyenne	89.4 *(3.4)*	88.3 *(4.8)*	89.7 *(4.2)*
écart-type	6.7	3.1	9.8

Puissance de détection de l'association

Conclusion

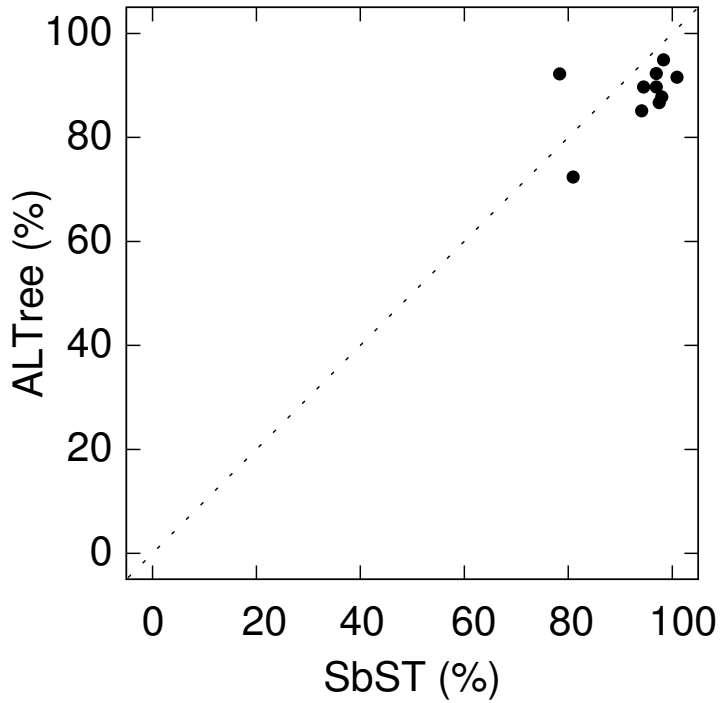
- L'utilisation de la phylogénie entraîne :
  - un léger gain de puissance par rapport au test haplotypique
  - mais pas de gain de puissance par rapport au test SbS
- Variabilité importante entre les jeux de données
  - peu liée au LD
  - facteurs liés à l'arbre (position site de susceptibilité, effectifs dans les branches...)

Résultats (1 site de susceptibilité simulé)

	ALTree	HT	SbST
1 site inclus parmi les marqueurs étudiés			
moyenne	92.9 *(3.5)*	91.2 *(4.8)*	99.35 *(4.4)*
écart-type	5.63	0.41	0.13
1 site non inclus parmi les marqueurs étudiés			
moyenne	89.4 *(3.4)*	88.3 *(4.8)*	89.7 *(4.2)*
écart-type	6.7	3.1	9.8

Efficacité d'identification des sites de susceptibilité

1 locus à risque

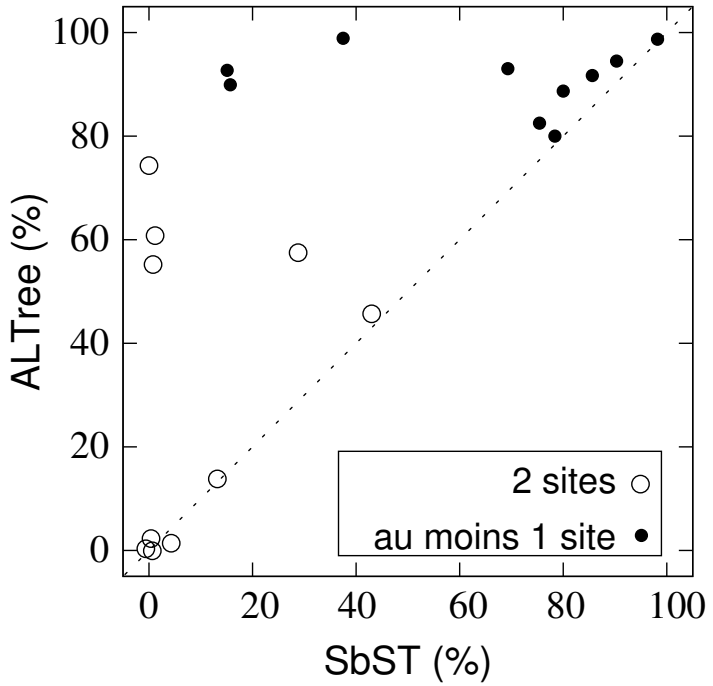


fréquence : 0.2

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

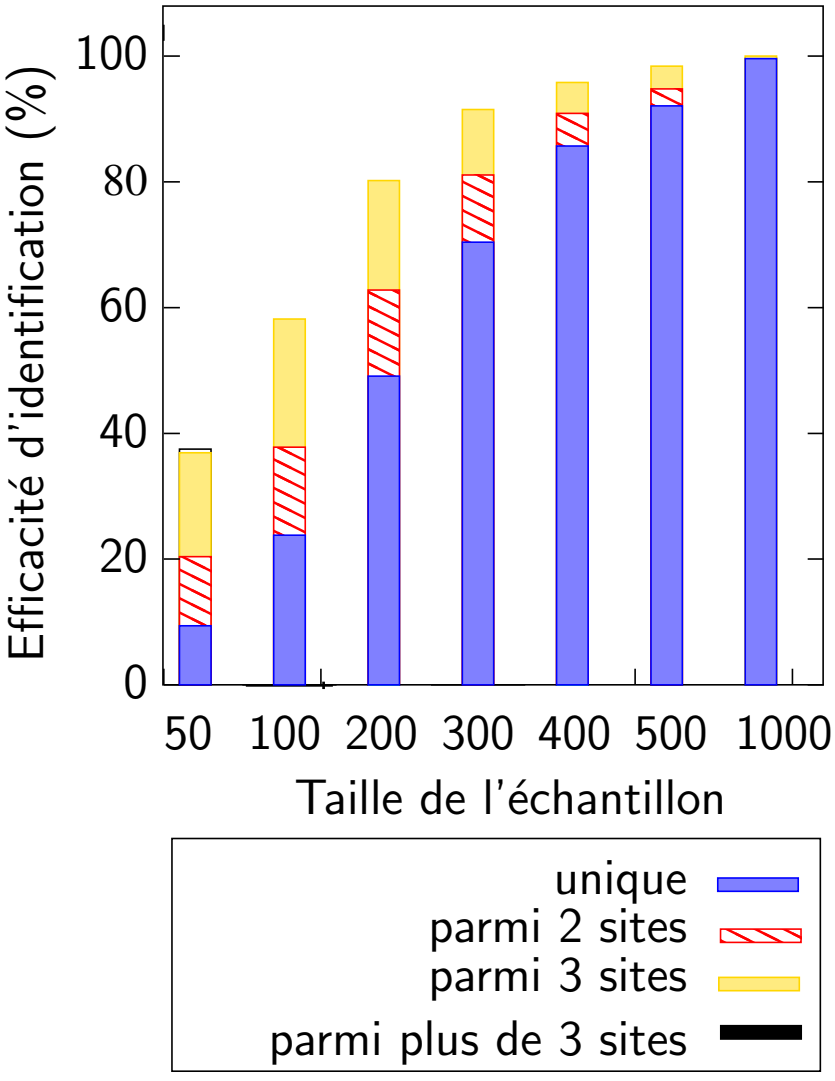
Les facteurs influençant la méthode

Les différents facteurs étudiés

- La taille de l'échantillon
- Le modèle de la maladie
  - La fréquence de l'allèle à risque
  - La pénétrance de la maladie
- Les facteurs liés à la reconstruction de l'arbre
  - La méthode de reconstruction de l'arbre
  - La méthode d'optimisation des caractères
  - L'enracinement

Étude de l'efficacité de localisation

Impact de la taille de l'échantillon



Conclusion

Lorsque la taille de l'échantillon augmente

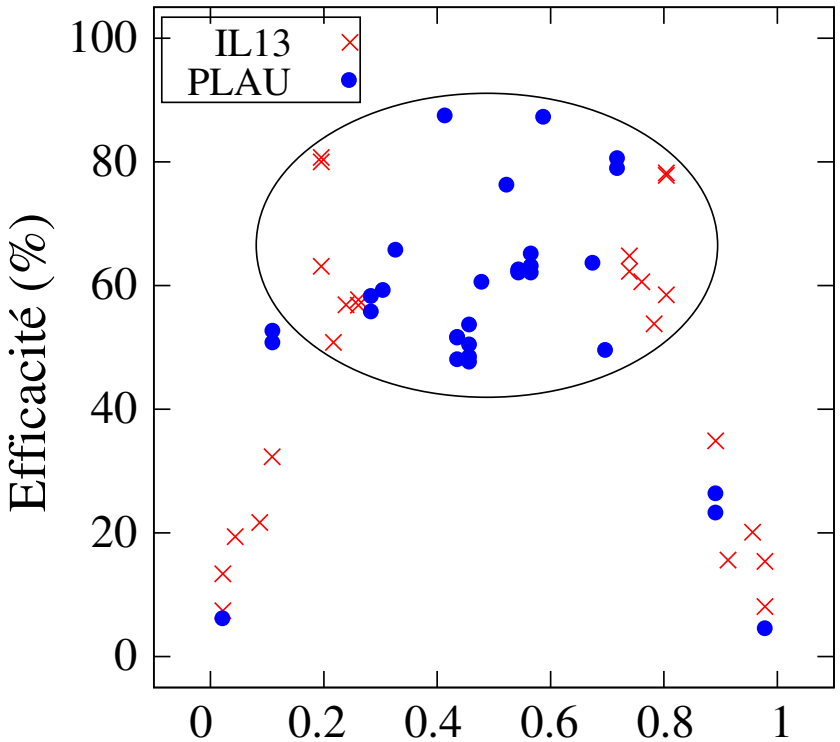
- Comme attendu, l'efficacité augmente
- L'identification du site de susceptibilité est moins ambiguë

Modèle simulé (gène PLAU) :

- freq de l'allèle à risque : 0.3
- $P_{\text{hom non porteur}}$: 0.01
- $P_{\text{hétérozygotes}}$: 0.02
- $P_{\text{hom porteur}}$: 0.1

Impact du modèle de la maladie

Exemple de la fréquence de l'allèle de susceptibilité



échantillon : 200 malades + 200 témoins
pénétrances : 0.03 (hom non porteur)
0.06 (hétérozygotes
0.30 (hom porteurs)

Conclusion

Dans nos conditions de simulation

- Efficacité maximum pour $0.2 < f < 0.8$
  - Pas de relation entre la fréquence et l'efficacité
- Efficacité plus faible pour des f extrêmes
  - malades non porteurs (f faible)
  - témoins porteurs (f élevée)

Plan de l'exposé

Données sur la maladie de Crohn (1)

Les données (fournies par J. P. Hugot)

- Maladie inflammatoire de l'intestin
- 232 familles génotypées pour 13 SNPs du gène CARD15
- Utilisation des haplotypes non transmis par les parents à leur enfant atteint comme témoins
- Reconstruction des haplotypes (logiciel GENEHUNTER)
- Restreindre les haplotypes à une zone **avec peu de recombinaison** (LD + Utilisation des travaux de Vermeire et al.)

$\Rightarrow$ Jeu de données formé de 33 haplotypes différents, constitués de 7 SNPs.

Données sur la maladie de Crohn (2)

L'analyse

- Recontruction phylogénétique par ML
  - Recontruction de l'arbre par PHYML
  - optimisation des caractères par PAML

  $\longrightarrow$ Analyse d'un seul arbre
- Enracinement sur l'haplotype majoritaire (test d'association)

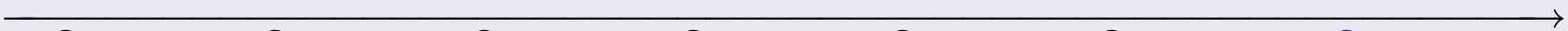
Résultats de l'analyse d'association

- Degré de signification obtenu par 10 000 permutations
- Test significatif : $p = 3 \times 10^{-4}$

Données sur la maladie de Crohn (3)

Résultats du test de localisation

Classement des 7 SNPs selon les indices de co-évolution V_i croissants



Analyse pas à pas (inspirée de Payami et al, 1989)

- Principe
  - Réaliser le test d'association
  - Identifier le locus responsable de l'effet le plus fort
  - Éliminer les haplotypes portant cet allèle
  - Réitérer jusqu'à ce qu'aucune association ne soit détectée
- Résultats
  - Identification de 3 SNPs : SNP 13, SNP 12 et SNP 8
  - SNPs précédemment identifiés par Hugot *et al.* (2001)

Plan de l'exposé

Conclusion

Développement d'une nouvelle méthode basée sur des phylogénies

- Test d'association + Identification des locus de susceptibilité
- Implémentation dans le logiciel ALTree (http://claire.bardel.free.fr/software.html)
- Méthode généralisable à l'analyse de données quantitatives et à la recherche de QTN (quantitative trait nucleotide)

Étude de ses propriétés

- Étude par simulation
- Pas vraiment de gain de puissance par rapport aux méthodes classiques d'association
- Permet de faire des hypothèses concernant les locus à risque

Application à des données

- Données sur la maladie de Crohn et la schizophrénie

Perspectives

Le problème de la recombinaison

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

Augmenter les fonctionalités de la méthode

- Prendre en compte l'incertitude liée à la reconstruction haplotypique
- En ML : prendre en compte la probabilité des états de caractères aux nœuds

Applications de la méthode à d'autres jeux de données

- Données qualitatives sur la schizophrénie : nouveaux marqueurs en cours de génotypage dans le gène DRD2
- Données quantitatives sur le taux de TAFI (Thrombine-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

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