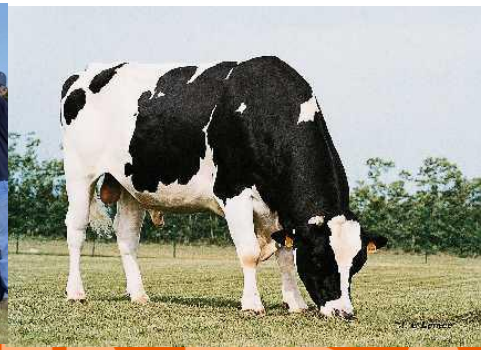
Haplotype Inference in Complex Pedigree

B. Kirkpatrick, E. Halperin and R.M. Karp



ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Objectifs

- Utiliser l'information haplotypique pour imputer les génotypes manquants dans les études d'association pan-génomiques
  - Mise en évidence de nouveaux polymorphismes d'intérêt
  - Gain de puissance
  - Mais… l'utilisation de grand pedigree est souvent associée à un temps de calcul très long.
- Pour les analyses de pedigree, un algorithme s'est montré particulièrement adéquat
  - Le « blocked Gibbs sampling »

Objectifs

- Un cas particulier du problème de reconstruction d'haplotype dans un pedigree est traité: Les pedigree complexes à lignées multiples
  - Si on s'intéresse aux régions du génome suffisamment liées pour qu'on puisse négliger la recombinaison lors de la méiose :
    - Si il y a peu de chance qu'une recombinaison ancestrale ait eu lieu au sein des haplotypes fondateurs alors le modèle de coalescence est utilisé (Gusfield, 2002)
    - Si une recombinaison ancestrale est détectée, alors elle est autorisée et l'haplotype fondateur n'est pas restreint au modèle de coalescence

Gibbs sampling (Geman and Geman, 1984)

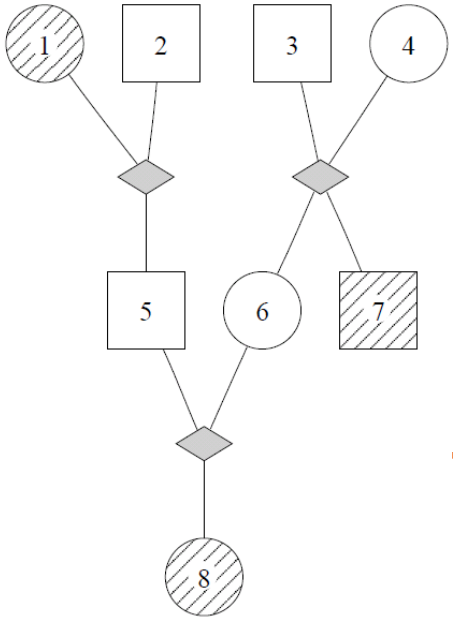
- Dans le cadre de la reconstruction d'haplotype, cet algorithme peut être décomposé en 3 étapes :
  1. Choisir un individu i au hasard dans l'échantillon
  2. Réestimer $H_i^{(t+1)}$ à partir de $pr(H_i \mid g, H_{-i}^{(t)})$, avec H_{-i} correspondant au set d'haplotypes en excluant l'individu i

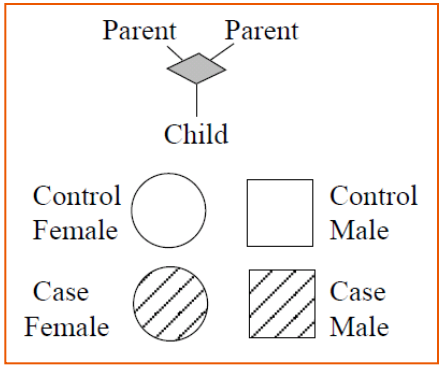
     $pr(H_i \mid g, H_{-i}^{(t)})$
  3. $H_j^{(t+1)} = H_j^{(t)}$ pour $j=1,\ldots,i-1,i+1,\ldots,n$.
- Cet algorithme suit une chaine de markov (MCMC)
- L'enjeu est de calculer correctement $pr(H_i \mid g, H_{-i}^{(t)})$
  - Voir travaux de Stephens, Nui et al.,

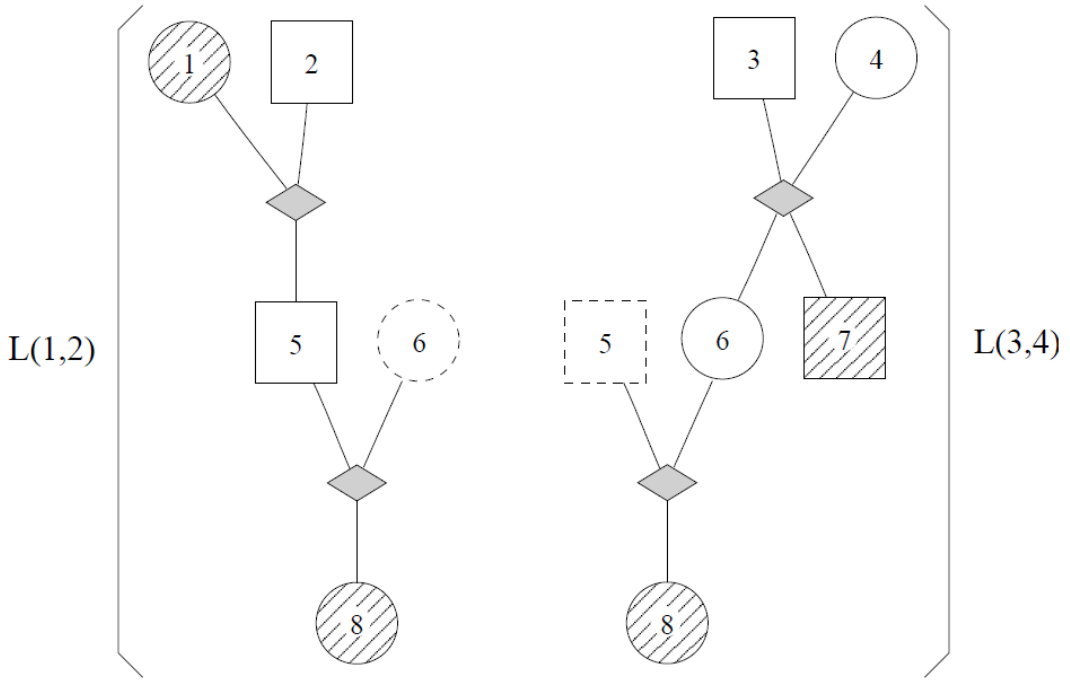
Blocked gibbs sampling (Jensen, 1997)

- Le « Blocked Gibbs sampling » est une généralisation du gibbs sampling où on travaille par groupe d'individus
  - La distribution des haplotypes est calculée intragroupe
  - L'objectif est d'inférer les états haplotypiques du pedigree complet à partir de la distribution a posteriori des haplotypes, des génotypes et de l'information sur les haplotypes partagés entre groupes d'individus
- Dans ce papier, les groupes d'individus sont définis en réalisant une décomposition par lignée
  - Une lignée correspond à un DAG (direct acyclic graph).
    La lignée est définie comme étant le plus petit sous pedigree intégrant les fondateurs de la lignée et leurs descendants
  - Les lignées ne sont pas nécessairement disjointes les unes des autres

Décomposition par lignée

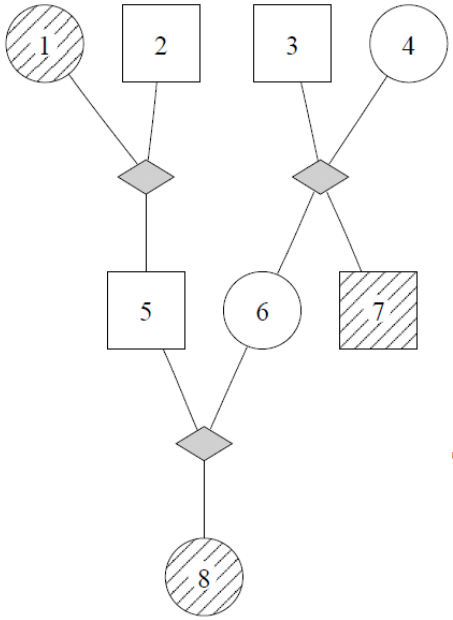


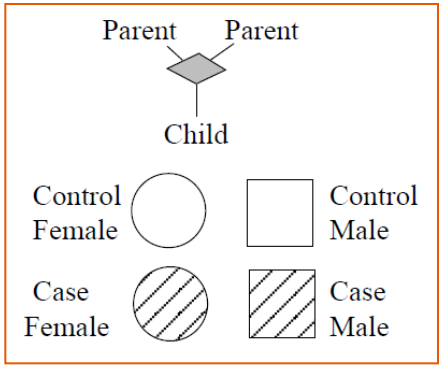


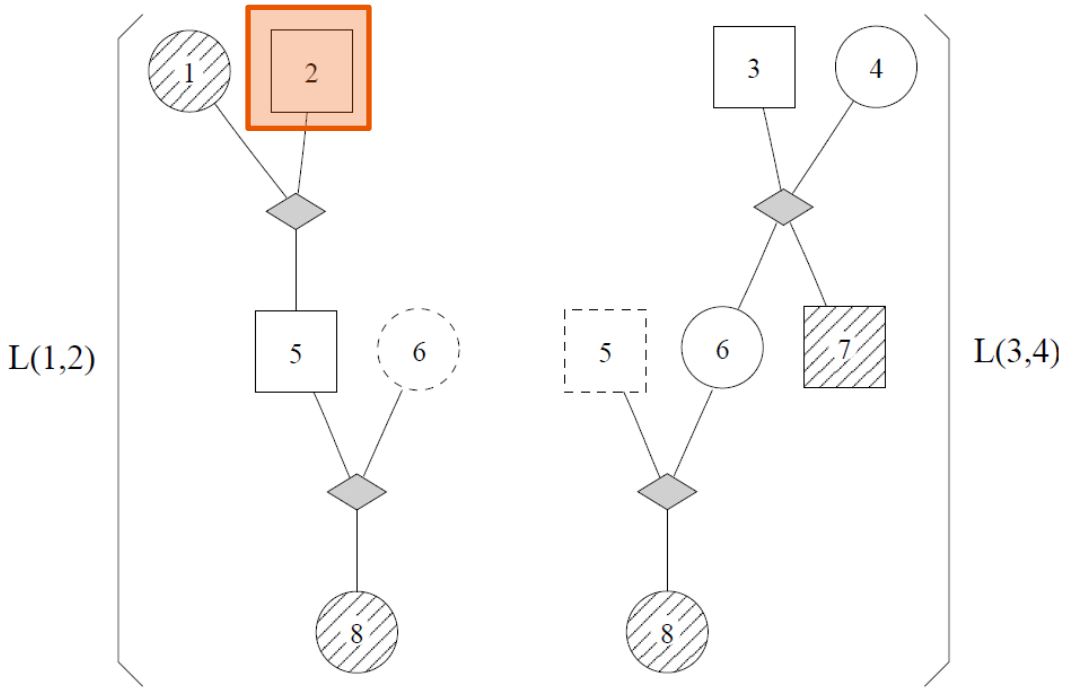


- Fondateurs
- Fondateurs qui ont des descendants dans d'autres lignées
- Descendants
- Parents qui ont des descendants dans d'autres lignées

Décomposition par lignée

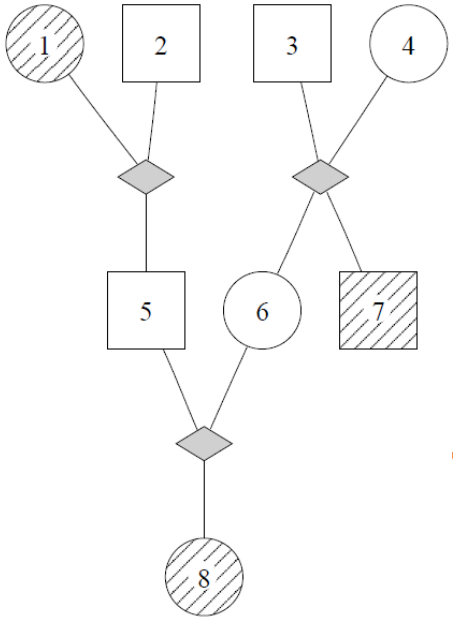


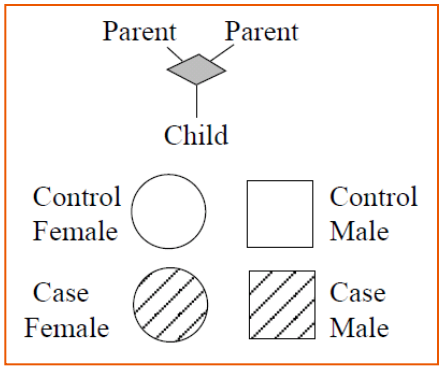


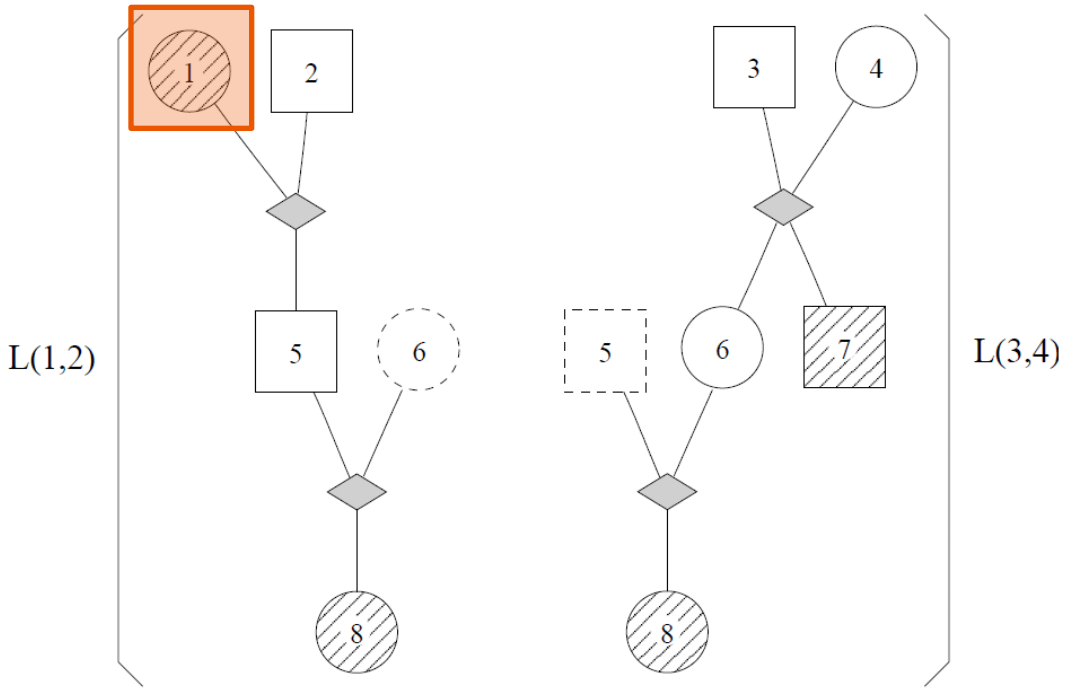


- Fondateurs
- Fondateurs qui ont des descendants dans d'autres lignées
- Descendants
- Parents qui ont des descendants dans d'autres lignées

Décomposition par lignée

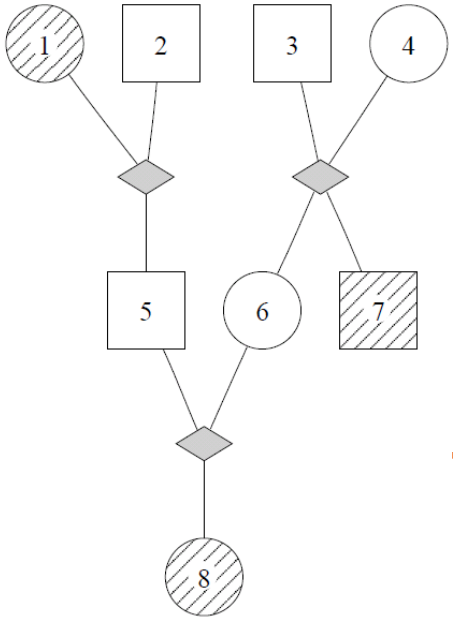


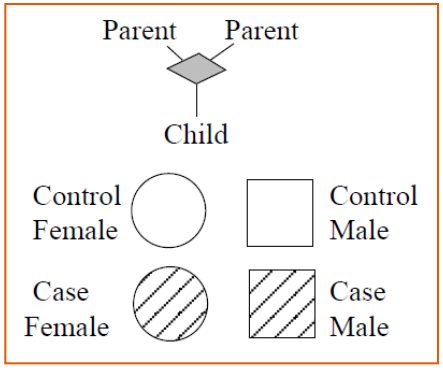


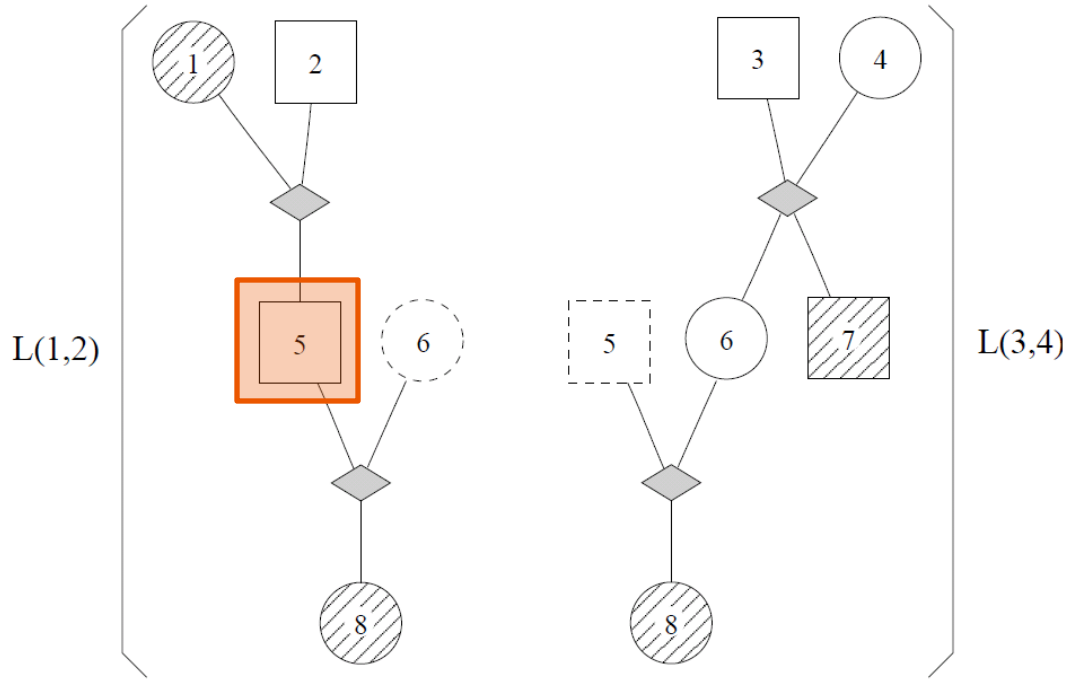


- Fondateurs
- Fondateurs qui ont des descendants dans d'autres lignées
- Descendants
- Parents qui ont des descendants dans d'autres lignées

Décomposition par lignée

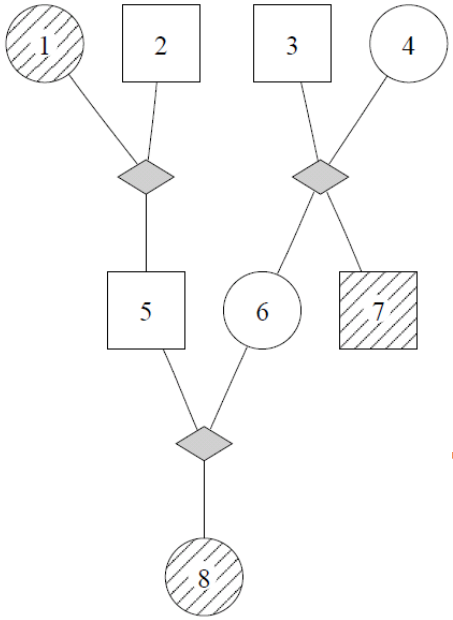


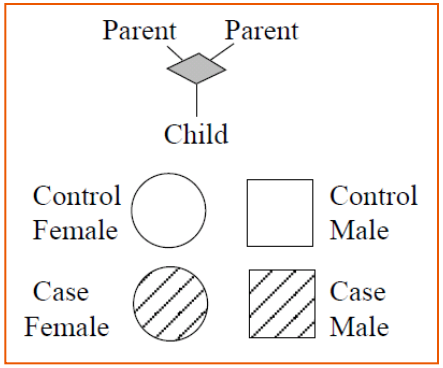


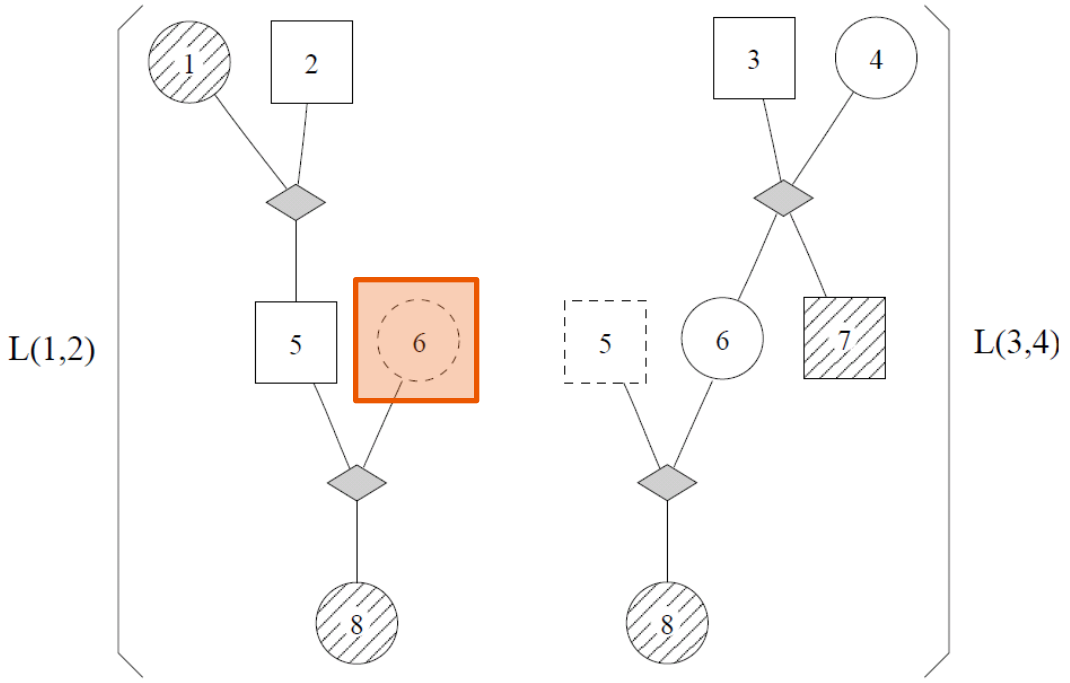


- Fondateurs
- Fondateurs qui ont des descendants dans d'autres lignées
- Descendants
- Parents qui ont des descendants dans d'autres lignées

Décomposition par lignée







- Fondateurs
- Fondateurs qui ont des descendants dans d'autres lignées
- Descendants
- Parents qui ont des descendants dans d'autres lignées

Blocked gibbs sampling

- Pour définir un état haplotypique, chaque lignée est traitée individuellement, la distribution des haplotypes est calculée chez les fondateurs en prenant en compte
  - Les données de génotypage
  - Les hypothèses liées aux fondateurs ou aux parents n'appartenant pas à la lignée
- Le calcul se fait en 3 étapes
  1. Trouver un état haplotypique cohérent sans recombinaison pour le pedigree (si le modèle de coalescence fonctionne, on le choisit, sinon on autorise une recombinaison ancestrale)
  2. Le pedigree est décomposé en lignées
  3. Sur chacune des lignées, on estime la distribution des haplotypes des fondateurs en fonction de l'information de génotypage de la lignée et des états haplotypiques des parents ayant des descendants dans d'autres lignées puis, assigner les haplotypes aux descendants de la lignée en fonction de ces distributions
- La longueur de l'haplotype est défini par l'utilisateur

Simulations

- Simulation d'un pedigrees et des haplotypes dont il est composé
- Pour chaque pedigree, les fondateurs sont générés à partir du chromosome 1 des individus de la population caucasienne de HAPMAP
- Les non fondateurs sont générés avec un taux de recombinaison fonction de la distance physique (loi de Poisson)
    - Moyenne de 2 recombinaisons sur le chromosome 1
- Les pedigrees ont une structure fixe
- Pour chaque pedigree, le nombre d'individus génotypés est fixé

Simulations: Les scénarios

- L1: 10 copies de familles de 20 individus avec 1 lignée et 13 individus génotypés
  1000 blocs de 3 SNPs avec une distance entre SNP moyenne de 11kpdb
- S1: une famille de 10 lignées de 59 individus dont 24 sont génotypés
  1000 blocs de 3 SNPs
- M1: 5 copies de familles S1
- M2: 10 copies de familles de 10 individus avec 2 lignées et 5 individus géntoypés
  10000 blocs de 3 SNPs
- H1: 16 individus, 2 lignées avec demi-frères et 9 individus génotypés
  300 blocs de 5 SNPs
- R1: 60 copies de familles de 12 individus avec 6 individus génotypés
  blocs de 4 SNPs
- R2: 30 copies de familles nucléaires de 7 personnes avec 5 pleins frères et 5 enfants génotypés
  blocs de 6 SNPs

Autres méthodes

- Efficacité de la reconstruction haplotypique
  - Merlin (Chen et Abecassis, 2007)
  - Superlink (Fishelson et al., 2005)
  - Ces deux logiciels calculent un maximum de vraisemblance sur des structures de pedigrees similaires
  - Le taux de recombinaison et les fréquences alleliques des fondateurs sont fixés dans le modèle (ce qui n'est pas le cas de Phyloped)
- Puissance de détection de la maladie
  - MQLS (Thornton and McPeek, 2007), quasi likelihood association score
  - Test de liaison de Merlin, (Abecassis et al, 2002)
  - Test d'association de Merlin (Burdick et al, 2006)

Mesure de l'efficacité

- L'efficacité correspond au pourcentage d'haplotypes inférés qui coincide avec les haplotypes vrais qui ont été simulés
- Cette définition ne permet pas de prendre en compte l'origine parentale des haplotypes
- Si une famille nécessite des haplotypes recombinants, Phyloped ne sera pas capable de les traiter
- Dans le calcul de l'efficacité, les auteurs ont choisis de pénaliser leur méthode l'absence de prédiction = efficacité nulle

Résultats

			Perfect Prior		Uninformative Prior	
Pedigree	Method		Avg	Std-Dev	Avg	Std-Dev
L1	PhyloPed		**0.867**	0.030	**0.867**	0.030
	Merlin		0.855	0.018	0.857	0.018
	Superlink		0.836	0.034	0.819	0.023
S1	PhyloPed		**0.809**	0.065	**0.809**	0.065
	Superlink		0.796	0.064	0.642	0.066
M1	PhyloPed		**0.808**	0.060	**0.808**	0.060
	Superlink		0.795	0.058	0.636	0.058
H1	PhyloPed		**0.816**	0.161	**0.816**	0.161
	Merlin		0.750	0.138	0.761	0.124
	Superlink		0.799	0.116	0.717	0.148

Peu de diff

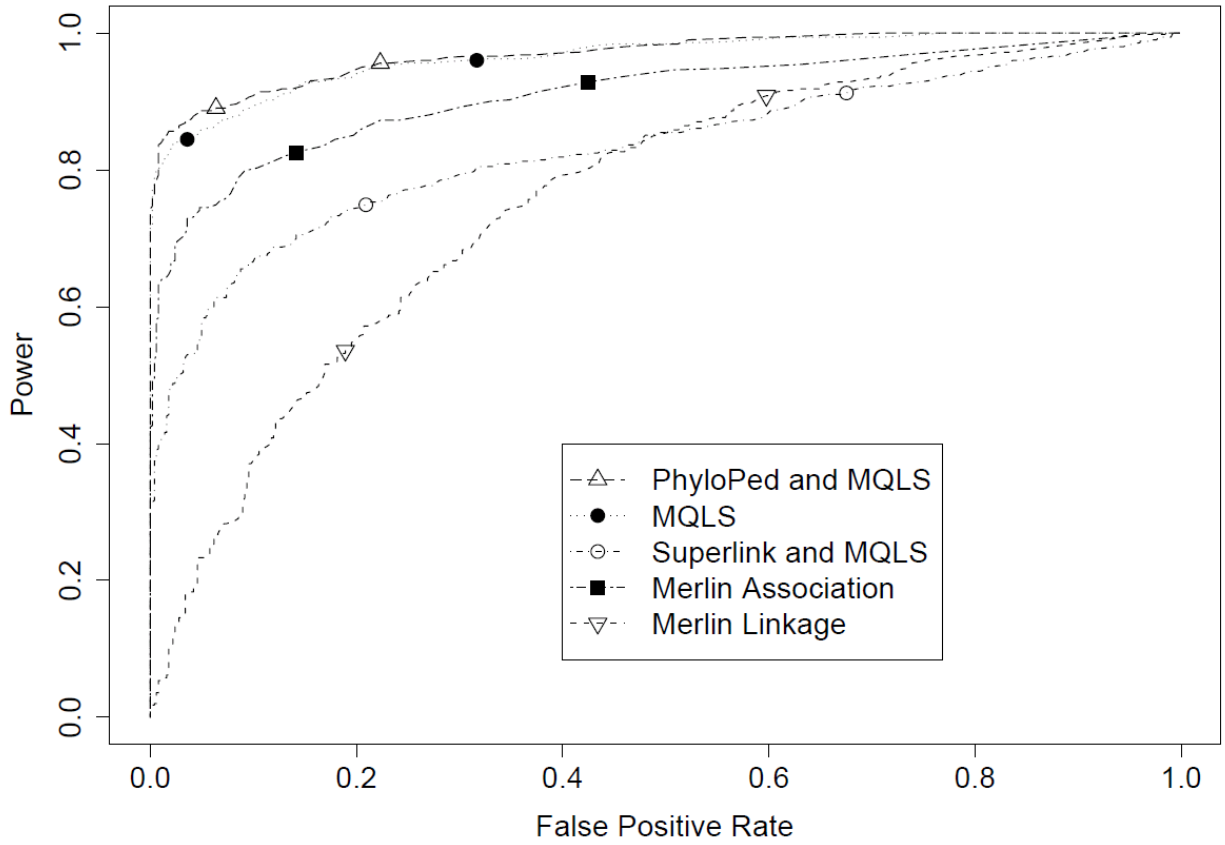
Phyloped meilleur

- L1: 10 copies de familles de 20 individus avec 1 lignée et 13 individus génotypés pedigree simples
- S1: une famille de 10 lignées de 59 individus dont 24 sont génotypés
- M1: 5 copies de familles S1
- H1: 16 individus, 2 lignées avec demi-frères et 9 individus génotypés

Pedigrees plus complexes

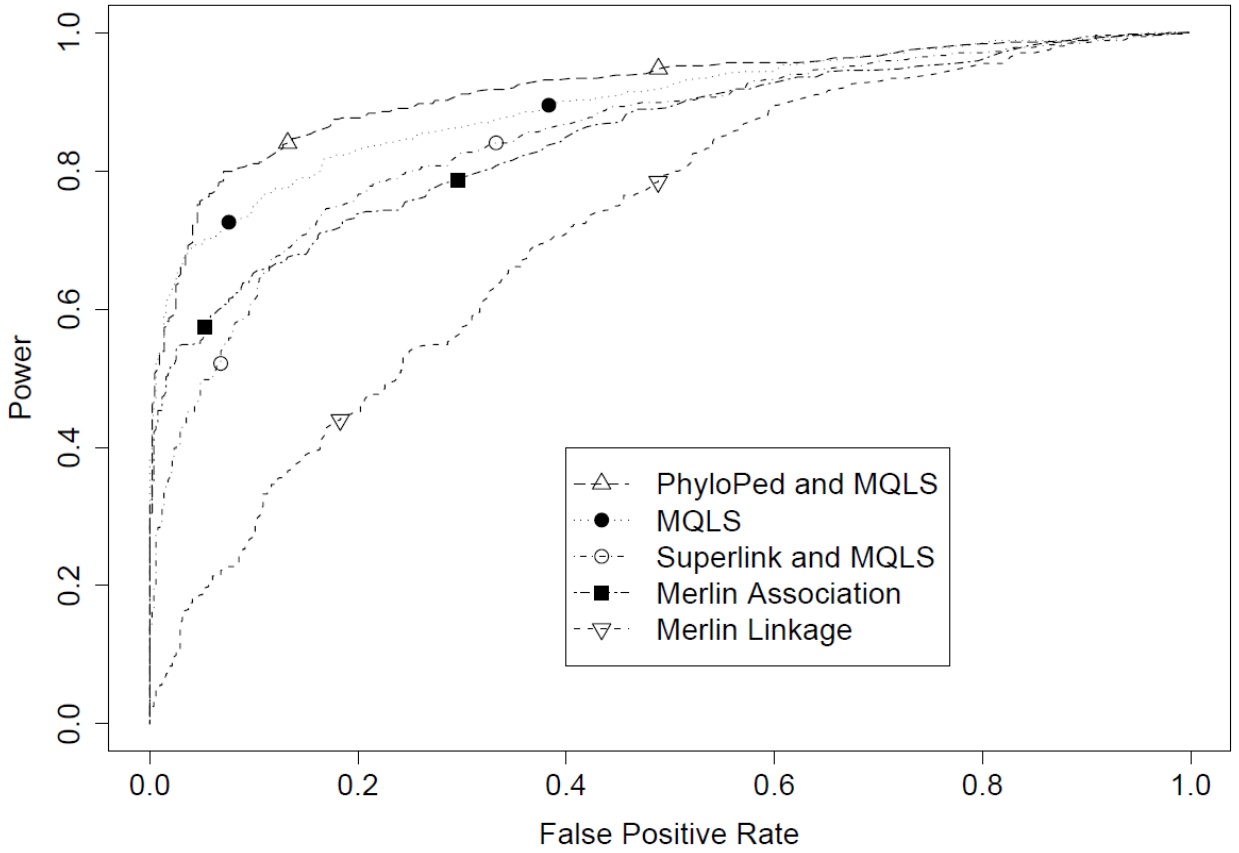
Résultats pour R1

60 copies de familles de 12 individus avec 6 individus génotypés, blocs de 4 SNPs



Résultats pour R2

30 copies de familles nucléaires de 7 personnes avec 5 pleins frères et 5 enfants génotypés
blocs de 6 SNPs



Discussion

- Phyloped
  - Haplotypes fondateur sont issus d'une phylogénie parfaite
  - Haplotypes sont transmis sans recombinaison dans le pedigree
- Cette méthode est efficace pour des petites régions avec une forte densité de SNP
- La combinaison Phyloped/MQLS est particulièrement efficace pour mener une étude d'association sur pedigree
- Aucune information a priori sur le taux de recombinaison ou sur les fréquences haplotypique chez les fondateurs n'est nécessaire à Phyloped
  - plus robuste
- Dans cette analyse, des haplotypes de 3 SNP ont été utilisés pour pouvoir comparer les résultats avec Merlin (limite de Merlin).
- Phyloped autorise des haplotypes de 5 SNP (résultats non montrés) avec de meilleurs résultats

Temps de calcul

Pedigree	Method		Avg Run-Time
L1	PhyloPed		0.150 s
	Merlin		0.166 s
	Superlink		0.072 s
S1	PhyloPed		0.612 s
	Superlink		0.041 s
M1	PhyloPed		18.2 s
	Superlink		0.156 s
H1	PhyloPed		3.69 s
	Merlin		0.088 s
	Superlink		17.5 s