

# Cartographie génétique par analyse de liaison

Brigitte Mangin

Septembre 2011

## 1 Introduction

## 2 Modélisation

- Les phénotypes binaires
- Les phénotypes continus

## 3 Gène “maladie”

- Le test du rapport de vraisemblance
- Quel seuil pour la liaison gène “maladie” - marqueur

## 4 Un QTL

- Le test du rapport de vraisemblance
- Quel seuil pour la détection d'un QTL ?
- L'approximation de la vraisemblance
- le LOD “support interval” du QTL

## 5 Plusieurs QTLs

## 6 Pour aller plus loin

- D'autres modélisations
- Pedigrees complexes
- Les logiciels

# Introduction

## Qu'est ce que l'analyse de liaison ?

⇒ C'est l'étude de la **transmission** allélique parmi des individus **apparentés**

## Que recherche-t'on ?

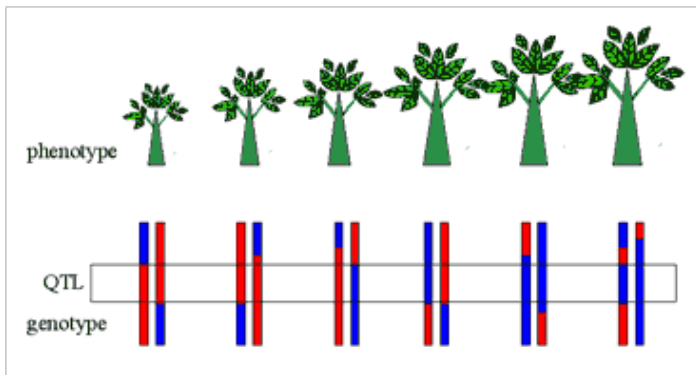
⇒ Des **gènes** causaux responsables d'une maladie, caractère binaire

ou

⇒ Des **QTLs**, c'est-à-dire des loci du génome qui expliquent la variabilité d'un caractère quantitatif.

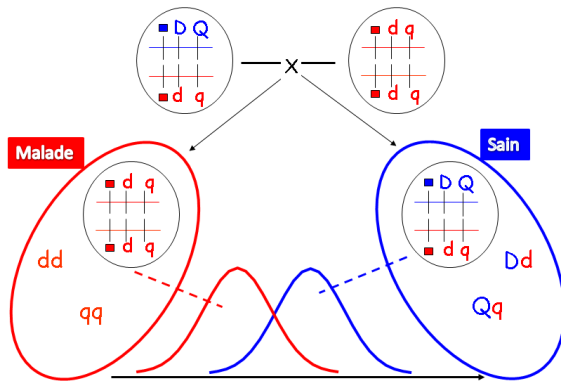
## Que cherche t-on à faire ?

⇒ lier le caractère binaire ou quantitatif (phénotype) à un endroit du génome (génotype au gène ou au QTL)



## Comment s'y prend-t'on ?

⇒ suit la transmission allélique à l'aide de marqueurs moléculaires



# Modélisation

## La modélisation pour quoi faire ?

⇒ “décrire” la relation entre le phénotype et le génotype

## En terme de probabilité “pénétrance”, phénotype binaire

$$f_{qq} = \text{Prob}(\text{malade} \mid qq)$$

$$f_{Qq} = \text{Prob}(\text{malade} \mid Qq)$$

$$f_{QQ} = \text{Prob}(\text{malade} \mid QQ)$$

Lien entre le phénotype et la maladie.

## En terme de densité de probabilité $\mathcal{L}$ , phénotype continu $Y$

$$f_{qq} = \mathcal{L}(Y \mid qq)$$

$$f_{Qq} = \mathcal{L}(Y \mid Qq)$$

$$f_{QQ} = \mathcal{L}(Y \mid QQ)$$

# Maladies simples dans des pedigrees simples

## Les deux modèles simples

$$f_{dd} = \text{Prob}(\text{malade} \mid qq)$$

1

pénétrance complète

$$f_{Dd} = \text{Prob}(\text{malade} \mid Qq)$$

=  $f_{dd}$ 

dominante

$$f_{DD} = \text{Prob}(\text{malade} \mid QQ)$$

0

pas de phénocopie

1

pénétrance complète

=  $f_{DD}$ 

récessive

0

pas de phénocopie

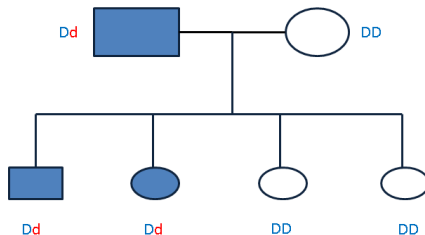
## Grâce au modèle postulé

⇒ Phénotypes "malade ou sain" et génotypes au locus/gène "maladie" sont en correspondance simple.

# Maladie simple dominante

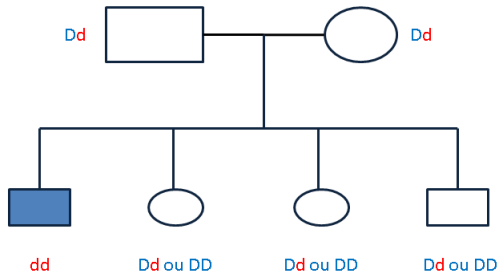
## Famille typique

Le pedigree des familles est simple car constitué de 2 générations : parents et enfants





# Famille typique pour une maladie simple récessive



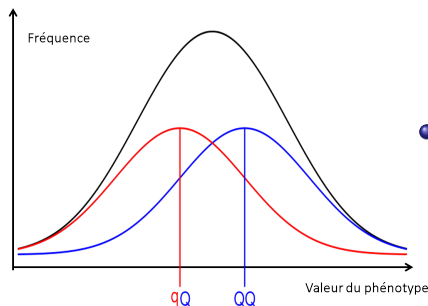
# Modèles QTL simples dans pedigrees simples

## Nous allons nous restreindre

- à une modélisation simple qui postule la présence d'un unique QTL
- à des descendance simples (back-cross et  $F_2$ )
- dans le cadre de la génétique mendélienne
  - back-cross : 1/2 qq , 1/2 Qq
  - $F_2$  : 1/4 QQ , 1/2 Qq , 1/4 qq

# Back-cross ou rétrocroisement

La distribution du phénotype (courbe noire)

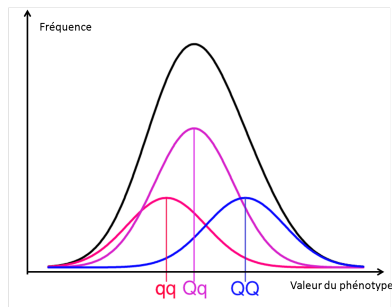


- est un mélange 1/2:1/2 entre les distributions du phénotype pour chaque génotype
- les distributions du phénotype pour chaque génotype ont la même variance, mais pas la même moyenne

- et sont Gaussiennes

# Descendance $F_2$

La distribution du phénotype (courbe noire)



- est un mélange  
1/4:1/2:1/4 entre les  
distributions du  
phénotype pour chaque  
génotype
- les distributions du  
phénotype pour chaque  
génotype ont la même  
variance, mais pas la  
même moyenne

- et sont Gaussiennes

## Les phénotypes continus

## Données

quantitative

 $Y_n$  $n = 1, \dots, N$  $n$  : individus

marqueurs

 $M_n^m$  $m = 1, \dots, M$  $m$  : marqueurs

## Paramètres du modèle

 $\sigma^2$ 

variance résiduelle

 $\mu_g$ effet moyen du génotype  
 $g$  au QTL

## Modèle de mélange

$$Y_n = \sum_g X_{n,g} \mu_g + \epsilon_n$$

$$X_{n,g} = \begin{cases} 1 & \text{si } n \text{ a le génotype } g \text{ au QTL} \\ 0 & \text{sinon} \end{cases}$$

## Les phénotypes continus

$$Y_n = \sum_g X_{n,g} \mu_g + \epsilon_n$$

## Back-cross

g=1 → génotype qq    ou    g=1 → génotype QQ  
g=2 → génotype Qq    g=2 → génotype Qq

Descendance  $F_2$ 

g=1 → génotype QQ  
g=2 → génotype Qq  
g=3 → génotype qq

# Les paramètres génétiques

## Back-cross

fond génétique  $\mu = \frac{\mu_{QQ} + \mu_{Qq}}{2}$   
 effet de substitution allélique  $\alpha = \mu_{QQ} - \mu_{Qq}$   
 ou  $\alpha = \mu_{qq} - \mu_{Qq}$

## Descendance F<sub>2</sub>

fond génétique  $\mu = \frac{\mu_{QQ} + 2\mu_{Qq} + \mu_{qq}}{4}$   
 effet d'additivité  $a = \frac{\mu_{QQ} - \mu_{qq}}{2}$   
 effet de dominance  $d = \mu_{Qq} - \frac{\mu_{QQ} + \mu_{qq}}{2}$

# et les marqueurs à quoi y servent ?

- à **connaître** le génotype au QTL
  - lorsque le QTL et le marqueur sont au même locus
- à **inférer** le génotype au QTL, "pseudo-marqueur"
  - en calculant la probabilité que  $X_{n,g} = 1$  sachant les marqueurs et la carte génétique
- ou à **imputer** le génotype au QTL, "pseudo-marquage"
  - en tirant aléatoirement le génotype au QTL avec les probabilités que  $X_{n,g} = 1$  sachant les marqueurs et la carte génétique



# Cas des gènes de “maladie simple”

## Cartographier le gène “maladie” c’est :

- pour chaque marqueur, se poser la question : “le gène est-il en liaison génétique avec le marqueur ?” en prenant un risque  $\Rightarrow$  **c’est donc une procédure de test**
  - Quelle est l’hypothèse nulle ?
  - Quelle est l’hypothèse alternative ?
  - Quelle statistique de test ?
- le localiser sur le génome en répétant le test pour tous les marqueurs  $\Rightarrow$  **c’est donc une procédure de tests multiples**
  - Quel est le seuil de rejet de l’hypothèse nulle ?

# Tester la liaison génétique avec le marqueur

## Ressemble au test de la liaison entre deux marqueurs

En effet, le phénotype des maladies simples peut être vu “mathématiquement parlant” comme un marqueur spécial.

Hypothèse nulle  $H_0 = \{ r_{M|m} = 1/2 \}$

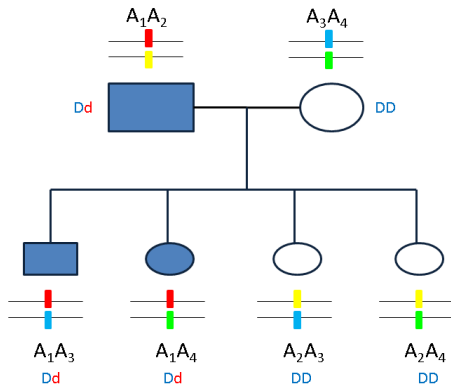
Hypothèse alternative  $H_1 = \{ r_{M|m} < 1/2 \}$

avec  $r_{M|m}$  le taux de recombinaison entre le marqueur  $M$  et le locus/gène “maladie”  $m$ .

# Maladie simple dominante

## Une difficulté

La phase du père est inconnu

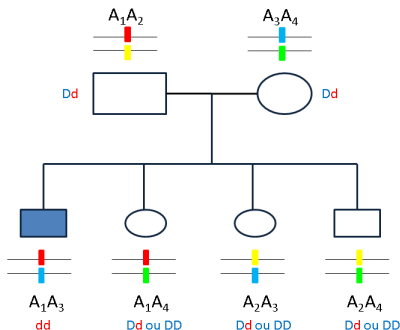


# Maladie simple récessive

## Les difficultés

Les phases des parents sont inconnues

Le génotype au gène "maladie" des enfants sains est imprécis



## Le rapport du maximum de vraisemblance

$$RV = \frac{V_{\mathcal{M}}(Y; r_{M|m}=1/2)}{\sup_{r_{M|m} < 1/2} V_{\mathcal{M}}(Y; r_{M|m})} \quad \text{toujours } \leq 1$$

$\mathcal{M}$                       génotypes aux marqueurs des enfants et des parents

$r_{M|m}$                     taux de recombinaison entre le marqueur  $M$  et le locus/gène "maladie"  $I^m$

$V_{\mathcal{M}}(Y; \dots)$         vraisemblance des phénotypes malades/sains  $Y$  conditionnelle à  $\mathcal{M}$

## Le LOD score

$$\text{LOD} = -\log_{10}(RV)$$

toujours  $\geq 0$

Quel seuil pour la liaison gène "maladie" - marqueur

# Seuil de rejet

**$\text{Prob}_{H_0}(\text{LOD} > \text{seuil}) \leq \text{risque}$**

Cette probabilité dépend :

- du modèle pour le gène/locus "maladie"
- de la connaissance ou non des phases des parents

**Pour un marqueur comme pour plusieurs marqueurs**

L'usage est de prendre un seuil pour le LOD score de 3.

La vraisemblance sous  $H_1$  est alors 1000 fois plus probable que la vraisemblance sous  $H_0$ .

$\text{LOD} > 3$                 rejette  $H_0$

$\text{LOD} < 3$     on ne rejette pas  $H_0$

# Cas des QTL

## Détecter un QTL c'est :

- prendre un risque et affirmer qu'il y a un QTL  $\Rightarrow$  **c'est donc une procédure de test**
  - Quelle est l'hypothèse nulle ?
  - Quelle est l'hypothèse alternative ?
  - Quelle statistique de test ?
- le localiser sur le génome en répétant le test tout le long du génome  $\Rightarrow$  **c'est donc une procédure de tests multiples**
  - Quel est le seuil de rejet de l'hypothèse nulle ?
- estimer les paramètres du modèle

# Tester l'effet du QTL

Hypothèse nulle  $H_0 = \{ \text{tous les génotypes au QTL ont la même moyenne} \}$

Hypothèse alternative  $H_1 = \{ \text{l'un au moins des génotypes au QTL a une moyenne différente des autres} \}$

## Pour le Back-cross

$$H_0 = \{ \mu_{QQ} = \mu_{Qq} \} \text{ versus } H_1 = \{ \mu_{QQ} \neq \mu_{Qq} \}$$

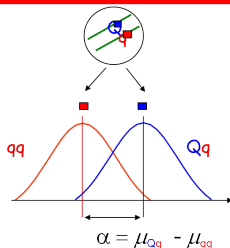
ou

$$H_0 = \{ \alpha = 0 \} \text{ versus } H_1 = \{ \alpha \neq 0 \}$$



# Le cas idéal : un marqueur sur le locus du QTL

## Modèle linéaire : comparaison de moyennes

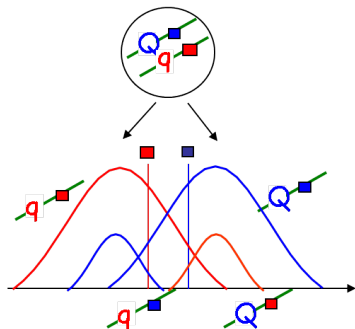

 $Y_n$ 

$$g = 1 \quad Y_{1k} = \mu + \frac{\alpha}{2} + \epsilon_{1k}$$

$$g = 2 \quad Y_{2k} = \mu - \frac{\alpha}{2} + \epsilon_{2k}$$

Source de variation	ddl	SCE	CM = $\frac{SCE}{ddl}$	F
QTL	1	$SCE_M = \sum_g n_g (Y_{g.} - Y_{..})^2$	CM <sub>M</sub>	$\frac{CM_M}{CM_R}$
résiduelle	N - 1	$SCE_R = \sum_{gk} (Y_{gk} - Y_{g.})^2$	CM <sub>R</sub>	

# Le cas non idéal



⇒ c'est la méthode appelée "l'interval mapping"

- Se contenter de faire la comparaison de moyennes entre les ■ et les ■ c'est perdre de la puissance.
- Les marqueurs flanquants le locus QTL vont être utilisés pour inférer le génotype aux QTL

# Utilisation de tous les marqueurs : "l'interval mapping"

Le test du rapport du maximum de vraisemblance :

*Lander et Botstein, Genetics, 1989*

$$RV = \frac{\sup_{\mu, \sigma^2} V_{\mathcal{M}}(Y; \mu, l, \alpha = 0, \sigma^2)}{\sup_{\mu, l, \alpha, \sigma^2} V_{\mathcal{M}}(Y; \mu, l, \alpha, \sigma^2)}$$

- $\mathcal{M}$  génotypes aux marqueurs et **carte génétique**
- $l$  position du QTL sur la carte, définie par  $r_{M_i Q}$  et  $r_{QM_{i+1}}$   
les taux de recombinaison entre le QTL  
et ses deux marqueurs flanquants
- $V_{\mathcal{M}}(.; \dots)$  vraisemblance conditionnelle à  $\mathcal{M}$

# La vraisemblance

$$V_{\mathcal{M}}(Y; \mu, l, \alpha, \sigma^2) = \prod_n V_{\mathcal{M}}(Y_n; \mu, l, \alpha, \sigma^2)$$

indépendance des observations

$$Y_n \begin{cases} \nearrow & g=1 & \text{Prob}(X'_{1,n} = 1 | \mathcal{M}) & = \text{Prob}(G'_n = \text{QQ} | \mathcal{M}) \\ \searrow & g=2 & \text{Prob}(X'_{2,n} = 1 | \mathcal{M}) & = \text{Prob}(G'_n = \text{Qq} | \mathcal{M}) \end{cases}$$

$\Rightarrow$  La loi de distribution de  $Y_n$  est un mélange de loi Gaussiennes (notée  $\mathcal{N}(\cdot, \cdot)$ )

$$Y_n \sim \text{Prob}(G'_n = \text{QQ} | \mathcal{M}) \mathcal{N}(\mu + \frac{\alpha}{2}, \sigma^2)$$

$$+ \text{Prob}(G'_n = \text{Qq} | \mathcal{M}) \mathcal{N}(\mu - \frac{\alpha}{2}, \sigma^2)$$

# le LOD score

## La courbe du LOD

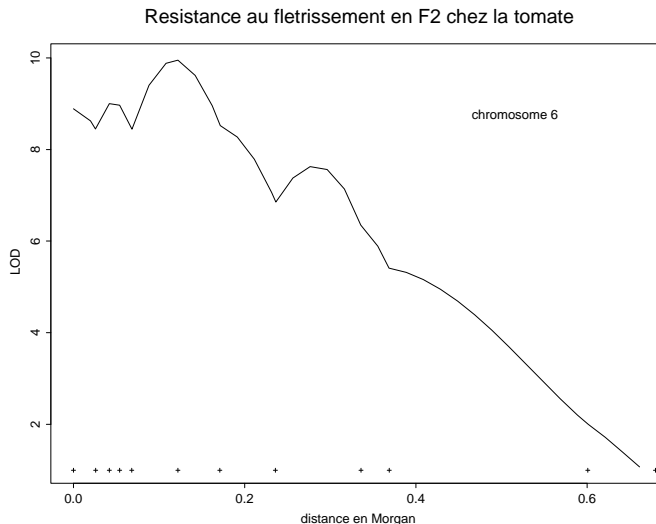
$$RV(l) = \frac{\sup_{\mu, \sigma^2} V_{\mathcal{M}}(Y; \mu, l, \alpha = 0, \sigma^2)}{\sup_{\mu, \alpha, \sigma^2} V_{\mathcal{M}}(Y; \mu, l, \alpha, \sigma^2)}$$

$$\text{LOD}(l) = -\log_{10}(RV(l))$$

## Le LOD test

$$\text{LOD} = -\log_{10}(RV) = \sup_l \text{LOD}(l)$$

# Un exemple de courbe de LOD



Quel seuil pour la détection d'un QTL ?

# Seuil de rejet d'un seul test en un seul locus

## Loi "asymptotique" sous $H_0$

$$\text{LOD}(l)/0.217 = -2 \ln(RV(l)) \sim \chi^2(\text{ddl}_{QTL})$$

Descendance	$\text{ddl}_{QTL}$
Backcross	1
$F_2$	2

## Pour un risque de première espèce donné,

$$\text{Prob}_{H_0}(\text{test}(l) > \text{seuil}) \leq 0.05$$

descendance	seuil en $-2 \ln(RV(l))$	seuil en $\text{LOD}(l)$
Backcross	3.84	0.83
$F_2$	5.99	1.30

Quel seuil pour la détection d'un QTL ?

# Seuil de rejet de l'ensemble des tests

## la question difficile des tests multiples non indépendants

Le seuil dépend

- du risque global choisi
- du nombre de groupes de liaison
- du nombre de marqueurs
- et de leur localisation
- du type de descendance



## Quel seuil pour la détection d'un QTL ?

$$\text{Prob}_{H_0}(\text{LOD} > \text{seuil}) \leq \text{risque}$$

⇒ formule analytique (étude asymptotique)

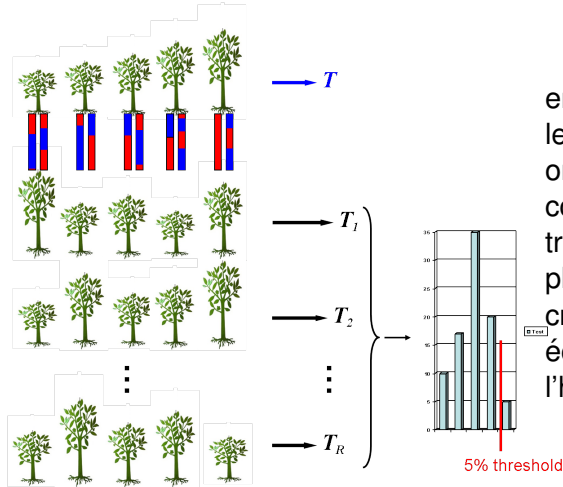
*Rebaï et al, Genetics, 1994*

⇒ simulation | par Monte-Carlo sous  $H_0$   
par permutation des  $Y_n$

*Churchill et Doerge, Genetics, 1994*

Quel seuil pour la détection d'un QTL ?

# Principe des permutations



en permutant les phénotypes, on casse la corrélation entre génotype et phénotype, on crée donc un échantillon sous l'hypothèse nulle

Quel seuil pour la détection d'un QTL ?

# Seuil pour le LOD test

		longueur du groupe de liaison : 1 M		
		intervalle entre les marqueurs		
type	risque	20 cM	10 cM	$\leq 1$ cM
BC	5 %	1.5	1.6	1.8
BC	1 %	2.2	2.3	2.6
F <sub>2</sub>	5 %	2.2	2.3	2.6
F <sub>2</sub>	1 %	2.9	3.1	3.4
		longueur du groupe de liaison : 2 M		
BC	5 %	1.8	1.9	2.1
BC	1 %	2.5	2.6	3.0
F <sub>2</sub>	5 %	2.4	2.6	3.0
F <sub>2</sub>	1%	3.2	3.4	3.8

# Modèle linéaire

## Au lieu du modèle de mélange

$$Y_n \sim \text{Prob}(G_n^I = \text{QQ} \mid \mathcal{M}) \mathcal{N}(\mu + \frac{\alpha}{2}, \sigma^2) \\ + \text{Prob}(G_n^I = \text{Qq} \mid \mathcal{M}) \mathcal{N}(\mu - \frac{\alpha}{2}, \sigma^2)$$

## Un modèle de régression sur marqueurs flanquants

*Haley et Knott, Genetics, 1991*

### Modèle linéaire pour un QTL en I

$$Y_n = \mu + \text{Prob}(G_n^I = \text{QQ} \mid \mathcal{M}) \frac{\alpha}{2} \\ - \text{Prob}(G_n^I = \text{Qq} \mid \mathcal{M}) \frac{\alpha}{2} + \epsilon_n$$

⇒ la statistique de test de Fisher en I est alors utilisée pour tester l'effet du QTL

## Équivalence asymptotique

Lorsqu'il y a beaucoup d'individus  
 un modèle de mélange et la statistique du LOD  
 ou  
 un modèle linéaire et la statistique de Fisher  
 cela détecte les mêmes QTL

## Mathématiquement parlant

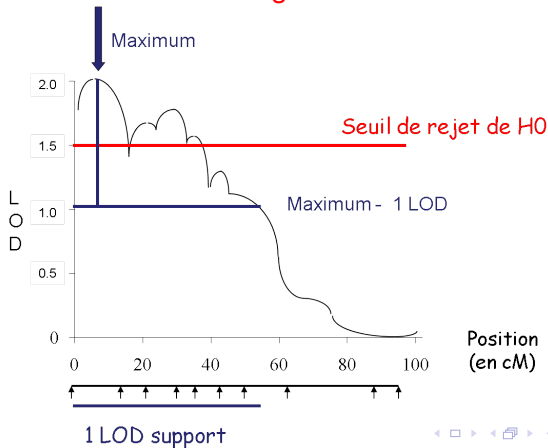
En tout point du génome, lorsque  $N \rightarrow \infty$   
 pour des QTL dont l'effet décroît avec  $\sqrt{N}$

$$\text{LOD}(l)/(2 * \ln(10)) = -2 \ln(RV(l)) \approx \text{ddl}_{\text{QTL}} F(l)$$

le LOD "support interval" du QTL

# Une région de confiance

Ce n'est pas un intervalle de confiance au sens statistique car on ne connaît pas la probabilité que le QTL soit dans de cette région



# Plusieurs QTL: utilisation des cofacteurs

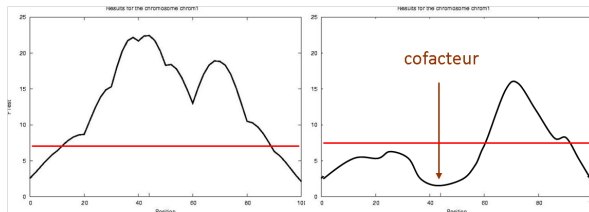
## Le cofacteur

- c'est un marqueur
- qui a été jugé explicatif
- par une méthode de choix de modèles (forward ou stepwise)
- dans le modèle linéaire

Il "absorbe" l'effet du QTL qui lui est proche.

# Objectif des cofacteurs

- contrôler les QTL des autres chromosomes pour gagner de la puissance par réduction de la variance de l'erreur résiduelle
- séparer les QTL d'un même chromosome

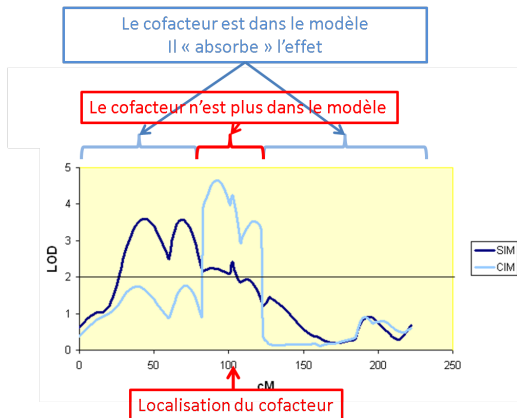




# Les différentes méthodes

- Composite Intervalle Mapping (CIM) = choix de cofacteurs puis détection de QTL dans un modèle de mélange
- iQTLm = méthode itérative de détection de QTL dans le modèle linéaire de régression sur marqueurs, les cofacteurs servant de point de départ de la méthode itérative
- le modèle bayésien avec un QTL dans chaque intervalle de marqueur et des lois a-priori pour les effets des QTL, le nombre et la localisation des QTL

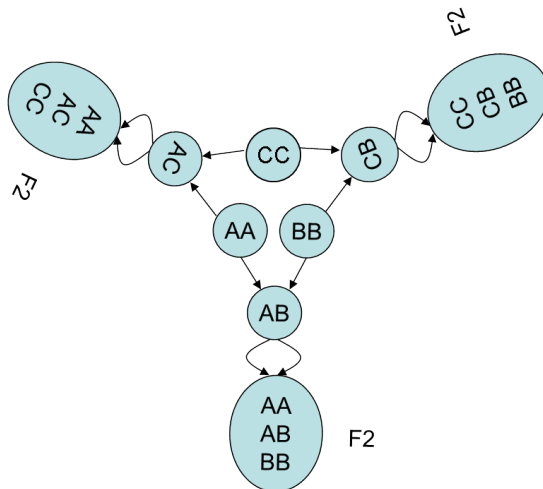
# Bien comprendre CIM



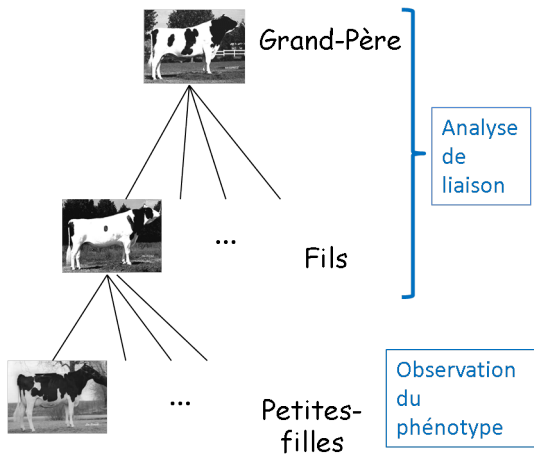
La courbe CIM dépend fortement de la place des cofacteurs. Il est primordiale de bien comprendre où ont été placés les cofacteurs.

- Effet du QTL aléatoire (animaux)
- Effet polygénique aléatoire (animaux) ou effet fond génétique fixe (plantes)
- Effet d'épistasie (interaction) entre QTLs
- Effet d'épistasie (interaction) entre un QTL et le fond génétique
- QTL en pléiotropie
- ....

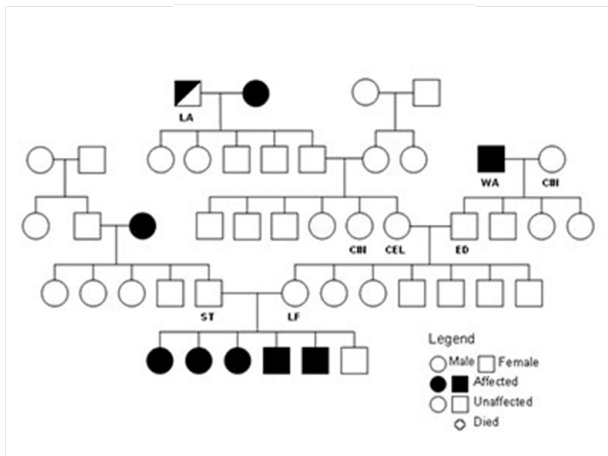
# Diallèle de lignées homozygotes



# Familles de demi-frères



# Pedigree humain



# Pour les QTL

- l'ancêtre : MAPMAKER/QTL
- le clé en main : QTL carthographier
- le payant : mapQTL
- le futur ? : R/qtl
- "les spécialisés"
  - pour les pedigrees et modèles animaux : QTLMap
  - pour plusieurs descendance : MCQTL