# Cartographie génétique

INRA, Thomas Schiex, Simon deGivry, Brigitte Mangin

Septembre 2016

# Plan

# Les cartes: s'orienter dans le génome

## Types

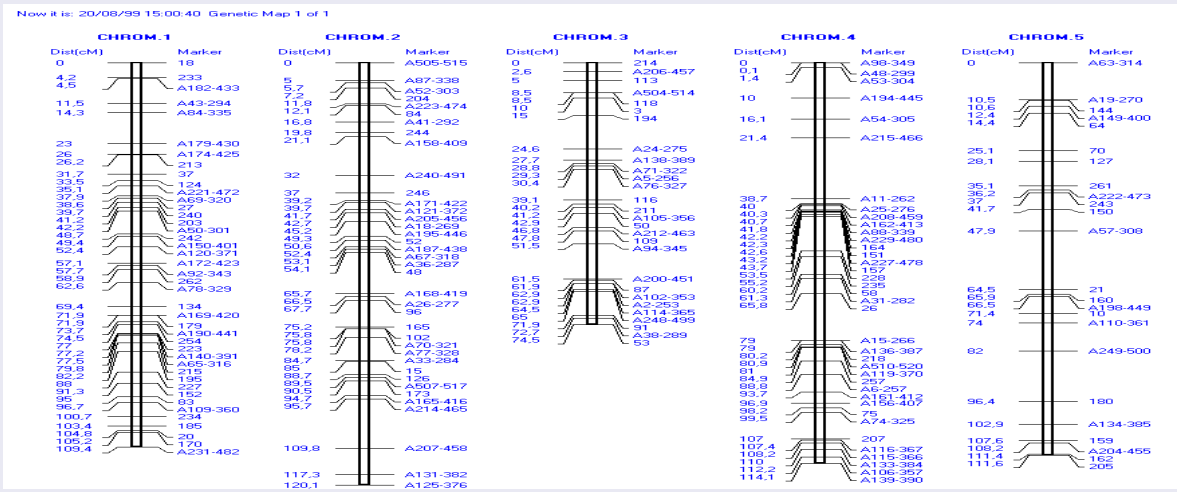
- **Cartes physiques** : distance réelle (Kb, Mb), à partir de fragments d'ADN. Résolution habituellement élevée.
- **Cartes d'hybrides irradiés** : Distance "statistique" liée à la cassure par irradiation, résolution intermédiaire.
- **Cartes génétiques** : s'appuie sur la recombinaison durant la méiose. Distance "statistique".

## Carte génétique/hybrides irradiés

Représentation d'un génome positionnant un ensemble de repères (marqueurs) dont on connaît les positions sur des groupes de liaison (chromosomes idéalement).

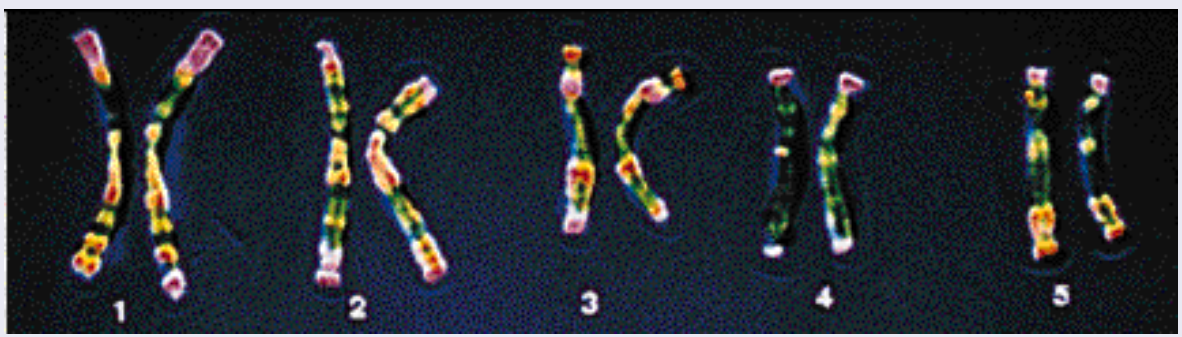
# Exemple

## Carte génétique




Groupes de liaison génétique

## Génome




Chromosomes

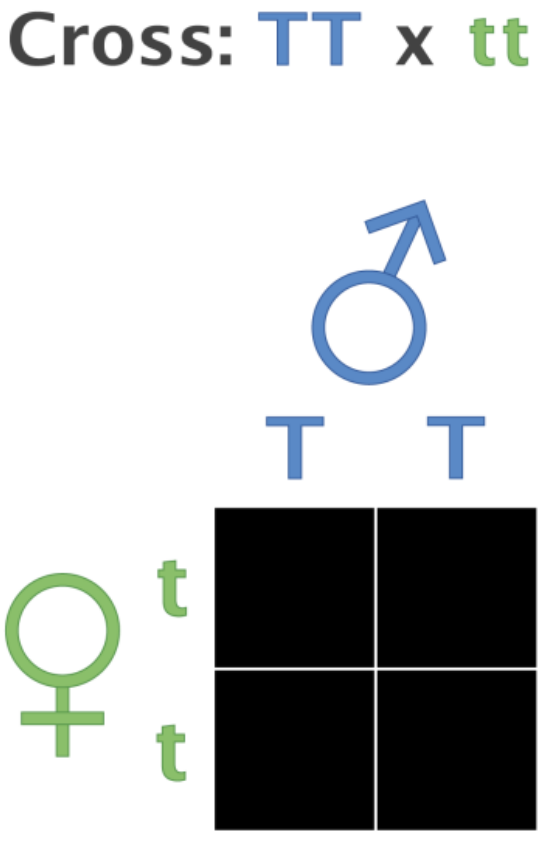
# Pourquoi

- Identifier les régions du génome influençant un caractère d'intérêt (maladie ou caractère quantitatif plus complexe)
- Positionner et identifier un gène (clonage positionnel)
- Comparer les génomes (étude de la synténie, évolution, transfert d'information)
- Faciliter la construction de cartes physiques, assemblage
- Étudier la méiose

# Les bases: lois de Mendel (modernes, diploïdes)

## Loi de ségrégation

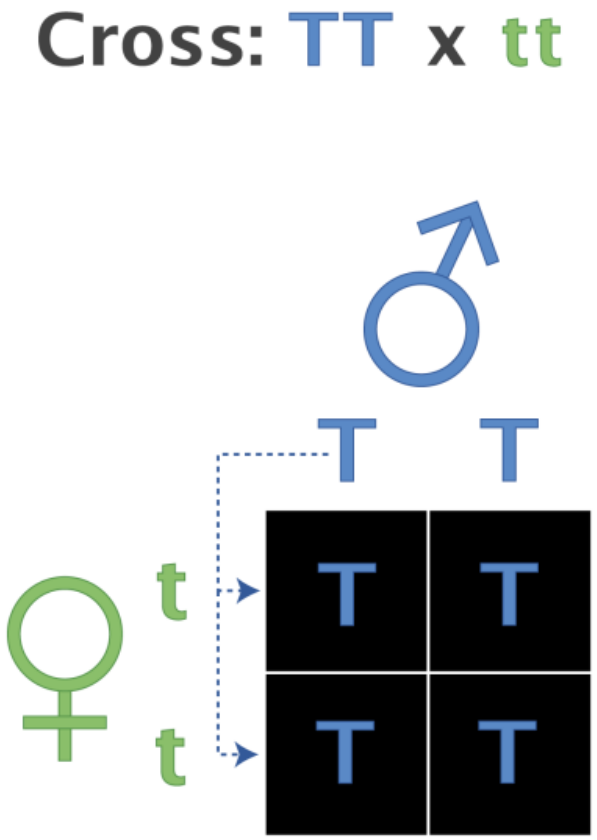
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles $Tt$).



# Les bases: lois de Mendel (modernes, diploïdes)

## Loi de ségrégation

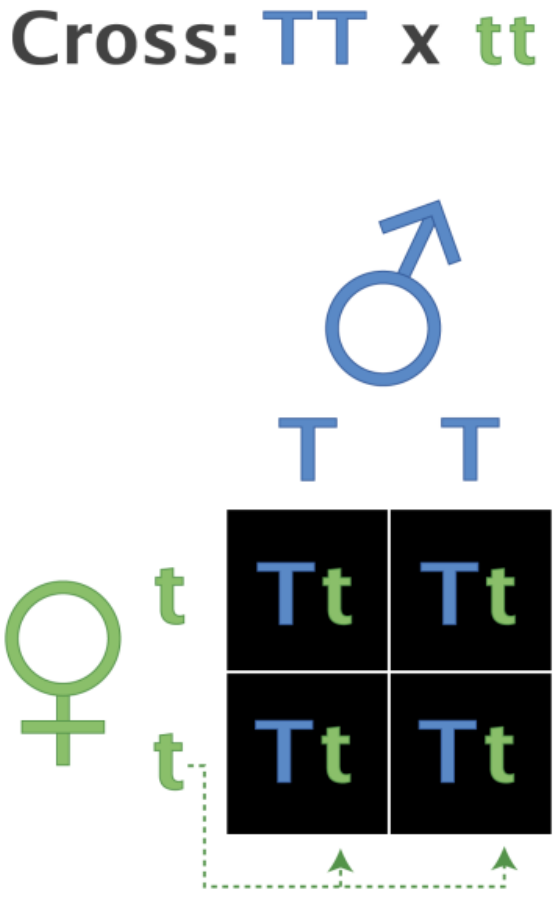
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles $Tt$).



# Les bases: lois de Mendel (modernes, diploïdes)

## Loi de ségrégation

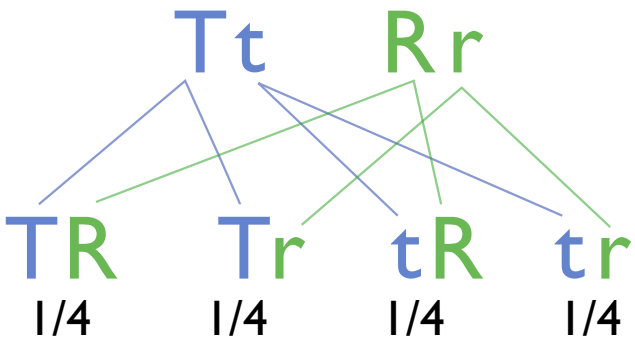
Un génome contient un ensemble de paires de gènes. Les paires ségrègent (se séparent) dans les gamètes, la moitié des gamètes portant un gène, l'autre moitié portant l'autre gène. Taille de plante (allèles $Tt$).



# Les bases: lois de Mendel (modernisées, diploïdes)

## Loi de ségrégation indépendante

L'assortiment de plusieurs gènes dans une cellule sexuelle se fait de façon indépendante entre les différents gènes. Taille $Tt$ et forme $Rr$ (ridé).

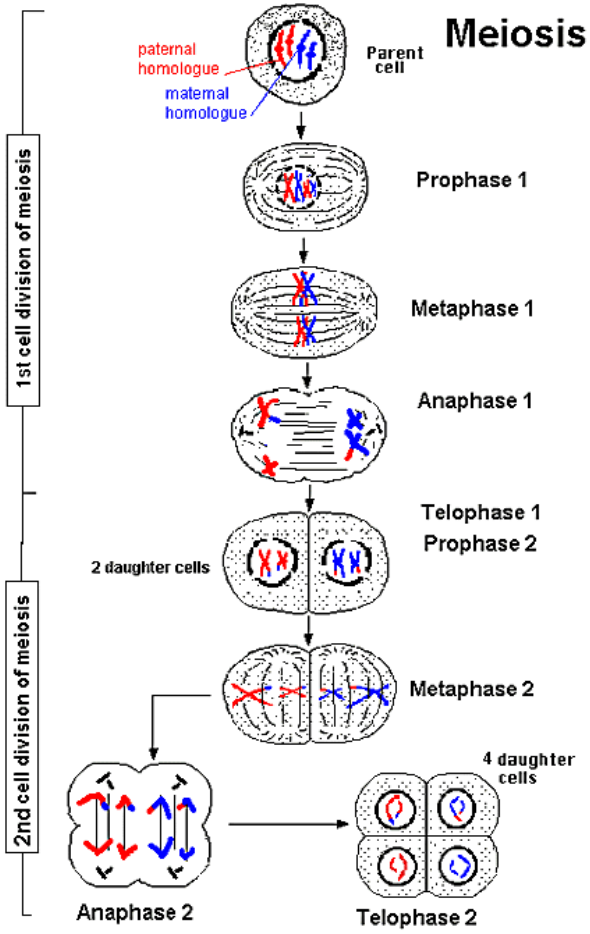


# Le principe historique de la cartographie

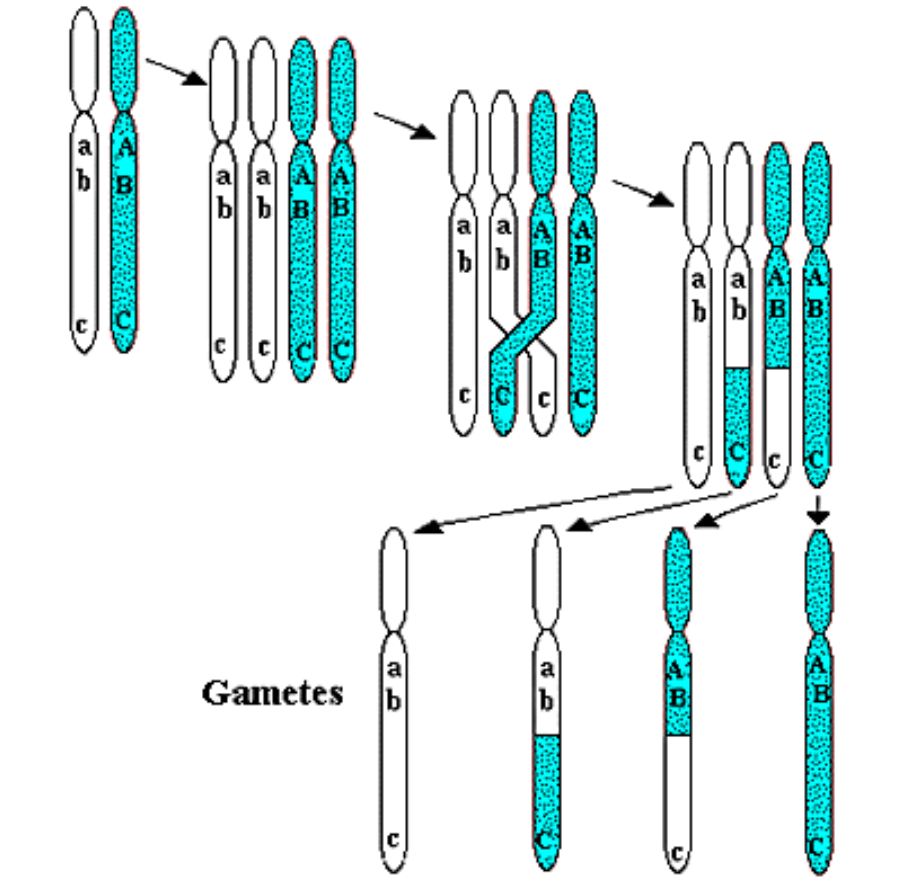
## Liaison génétique (Bateson 1905)

Pour certaines paires de gènes, la fréquence des combinaisons parentales dans les gamètes est supérieure à ce que l'on attend. On parle de **liaison génétique**.

Expliqué par Morgan (1911) par l'appartenance à un même chromosome et un éventuel chiasma durant la méiose (crossing-over).



# Un modèle de la méiose




Crossing-over and recombination during meiosis

50% de recombinants au plus.

# Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome

# Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles $Aa$).

# Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles $Aa$).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).

# Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles $Aa$).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portés par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).

# Bases

- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles $Aa$).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portés par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).
- **Génotype** : séquence des paires d'allèles (non ordonnées) portés par les chromosomes homologues ($\frac{A}{a}\ \frac{B}{b}$ par exemple).

# Bases

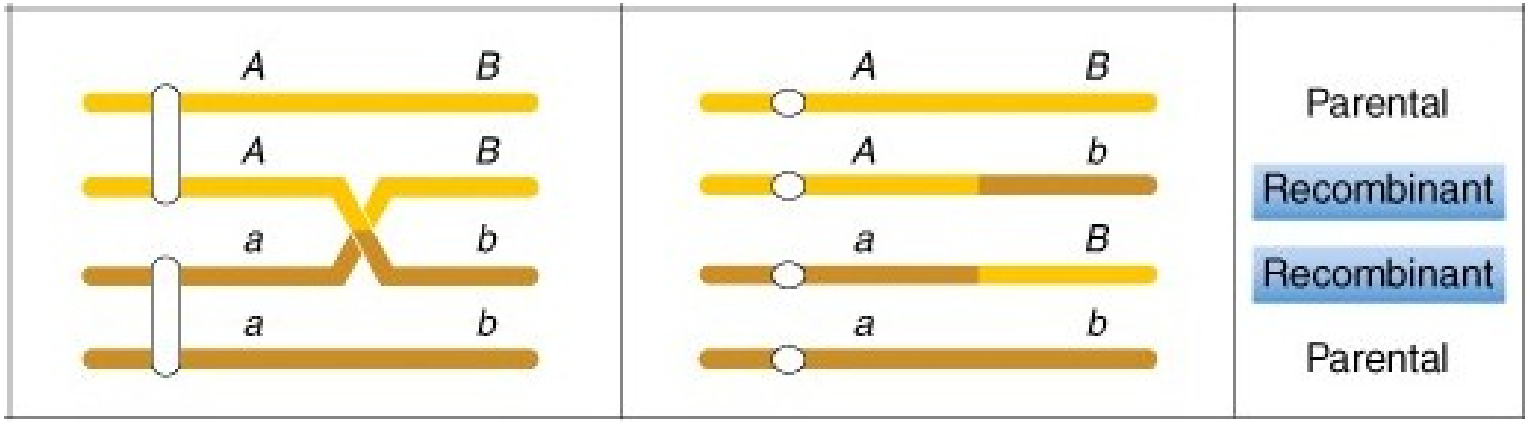
- **Loci, gènes, marqueurs** : $\mathcal{A}, \mathcal{B}$. Emplacement sur un chromosome
- **Polymorphisme** : présente au moins deux formes différentes (allèles $Aa$).
- **Homozygote** : paire d'allèles identiques ($\frac{A}{A}$ ou $\frac{a}{a}$), sinon hétérozygote ($\frac{A}{a}$).
- **Haplotype** : séquence des allèles portés par chacun des chromosomes ($\frac{Ab}{aB}$ par exemple).
- **Génotype** : séquence des paires d'allèles (non ordonnées) portés par les chromosomes homologues ($\frac{A}{a}\ \frac{B}{b}$ par exemple).
- **Phase**: information suffisante pour déterminer les deux haplotypes à partir du génotype.

# Recombinants et Non recombinants

## Marqueurs $\mathcal{A}, \mathcal{B}$

- une cellule diploïde portant les haplotypes $AB/ab$,
- on peut avoir les gamètes porteuses des haplotypes $AB, ab, Ab, aB$

Les deux premiers sont *parentaux* ou *non recombinants*. Les deux autres *recombinants* (nombre impair de cross-overs).



# Taux de recombinaison

**Taux de recombinaison $\leq \frac{1}{2}$**

Le taux de recombinaison $\rho_{\mathcal{AB}}$ entre les deux marqueurs $\mathcal{A}$ et $\mathcal{B}$ est la proportion de *recombinants*.

**Example**

Entre 3 gènes $\mathcal{Y}$ (yellow), $\mathcal{W}$ (white), $\mathcal{M}$ (miniature) de la drosophile, on observe $\rho_{\mathcal{Y},\mathcal{W}} = 1.3\%$, $\rho_{\mathcal{W},\mathcal{M}} = 32.6\%$ et $\rho_{\mathcal{Y},\mathcal{M}} = 33.8\%$. On peut penser que les marqueurs sont dans l'ordre $\mathcal{Y} - \mathcal{W} - \mathcal{M}$

Du fait des doubles crossing-overs, pour un ordre $\mathcal{Y} - \mathcal{W} - \mathcal{M}$ :

$$\rho_{\mathcal{Y},\mathcal{M}} < \rho_{\mathcal{Y},\mathcal{W}} + \rho_{\mathcal{W},\mathcal{M}} \quad \text{(non additif)}$$

# Distance génétique

## Définition

La distance génétique $d_{\mathcal{AB}}$ entre deux marqueurs $\mathcal{A}$ et $\mathcal{B}$ est le nombre moyen de *crossing-overs* entre les deux marqueurs par méiose.

## Propriétés

- Additif
- 1cM (centiMorgan) correspond à un crossover sur un haplotype pour 100 méioses.
- Les cross-overs ne sont pas facilement observables.

# Distance génétique et recombinaison

**Taux de recombinaison** : estimable à partir de données sur la descendance de parents bien choisis.

**Distance génétique** : s'appuie sur un modèle de la recombinaison.

## Interférence

Le taux de double recombinaison est habituellement inférieur à celui attendu sous hypothèse d'indépendance.

- $1.3\% \times 32.6\% = 0.43\%$ attendu pour la double recombinaison $\mathcal{Y} - \mathcal{W} - \mathcal{M}$.
- $0.045\%$ observé.

# Fonction de distance - map functions

Entre deux marqueurs. La première fonction de distance s'appuie sur un modèle de recombinaison simplifié (sans interférence, deux chromatides).

## Fonction de Haldane - sans interférence (1919)

$$\rho = \frac{1}{2}(1 - e^{-2d}) \qquad d = -\frac{1}{2}\log(1 - 2\rho)$$

Beaucoup d'autres fonctions pour l'interférence:

## Fonction de Kosambi - interférence (1944)

$$\rho = \frac{1}{2}\tanh(2d) \qquad d = \frac{1}{2}\tanh^{-1}(2\rho)$$

Pour de faibles distances/taux de recombinaison, $d \approx \rho$.

# Cartographie génétique

## Comment ?

1. Accumulation d'observations du génotype sur un ensemble de marqueurs et sur un bon nombre de méioses
   - Parents bien caractérisés (phase), hétérozygotie.
   - Observation sur la descendance
2. Reconstruire les distances et l'ordre des marqueurs.

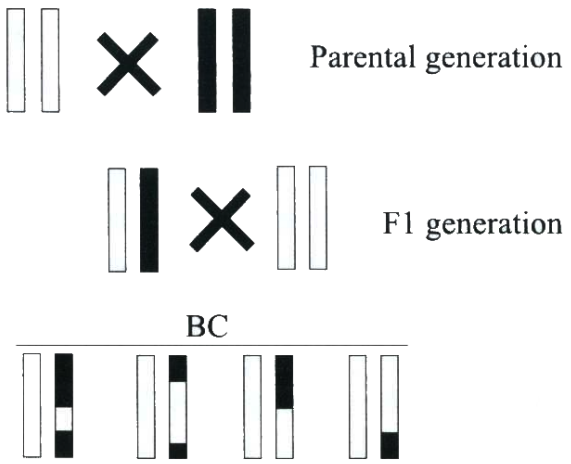
## Des situations variées

Taille de l'échantillon, temps de génération, mortalité des lignées, nombre de marqueurs, facilité des croisements (plantes, animaux, humain).
Observation de certains marqueurs/allèles parfois impossible (manquants).

# Back-cross: estimer $\rho$ entre 2 marqueurs

## Un individu peut avoir

- deux marqueurs homozygotes $(AA, BB)$ ou hétérozygotes $(Aa, Bb)$ : **non recombinant** $(NR)$.
- un hétérozygote, un homozygote $(Aa, BB$ ou $AA, Bb)$ : **recombinant** $(R)$.




## Vraisemblance - probabilité des données

Si $\rho$ est le taux de recombinaison (à estimer), la probabilité d'observer les données de typage $Data$ (la vraisemblance) est :

$$\text{Prob}(Data|\rho) = \rho^R(1-\rho)^{NR}$$

# Back-cross : estimer $\rho$ entre 2 marqueurs

## Vraisemblance - probabilité des données

$$\text{Prob}(Data|\rho) = \rho^R(1-\rho)^{NR}$$

## Maximum de vraisemblance

La valeur estimée $\hat{\rho}$ de $\rho$ choisie est celle qui maximise la probabilité d'observer les données (estimateur convergent).

Par un passage au logarithme et une étude de la dérivée on obtient :

$$\hat{\rho} = \frac{R}{R+NR}$$

# En pratique

- Individus non typés sur un marqueur. Données manquantes.
- On n'observe pas toujours les génotypes. Si un allèle $A$ est "dominant", $A$ est compatible avec $AA, Aa$ en back-cross.
- Erreurs de typages

La vraisemblance de données incomplètes est compliquée. De même que celle lorsque les marqueurs ne sont pas codominants, ou lorsque le pedigree est plus complexe que le back-cross.
Pour la maximiser on utilise des algorithmes d'optimisation dédiés (par exemple EM - Expectation Maximisation).

Dempster et al.,JRSS, 1977

# Nettoyage des données : distorsion

## Marqueur distordu

Allèle sur-représenté dans la descendance / à la fréquence attendue (gène lié à la reproduction/croissance, réarrangements ou problème d'échantillonage)

## Test de $\chi^2$ de Pearson $T_{\chi^2}$

Sous l'hypothèse nulle: $\{$ les données observées sont tirées de la distribution théorique attendue $\}$.

Pour un risque $\alpha = 0.05$, $\chi^2_{1ddl} = 3.84$

si $T_{\chi^2} > 3.84$ on rejette l'hypothèse de non distortion.

# Nettoyage des données : “marqueurs confondus”

## Jeux de données modernes

- typage de plusieurs dizaines de milliers de marqueurs
- distance minimale inter-marqueurs très faible
- pas de recombinaison/cassure obervée : même génotypes (ou génotypes compatibles avec les données manquantes).

$\Longrightarrow$ Supprimer ou fusionner des marqueurs

# Construction des groupes de liaison

## Groupes de liaison

Groupes de marqueurs qui appartiennent à un même chromosome (liés).

- Hypothèse 0: $\{$ 2 marqueurs $\mathcal{A}$ et $\mathcal{B}$ ont un assortiment indépendant (non liés, taux de recombinaison de $\frac{1}{2}$) $\}$.
- Hypothèse 1: $\{$ les 2 marqueurs $\mathcal{A}$ et $\mathcal{B}$ sont liés (taux de recombinaison $< \frac{1}{2}$) $\}$.

## $LOD$ score - 2 marqueurs

On utilise le $LOD$ score pour tester la liaison.

$$LOD = -\log_{10}\left(\frac{\text{maximum de la vraisemblance si } \rho{=}1/2}{\text{maximum de la vraisemblance}}\right)$$

Tradition : $LOD > 3$ utilisé pour conclure à la liaison.

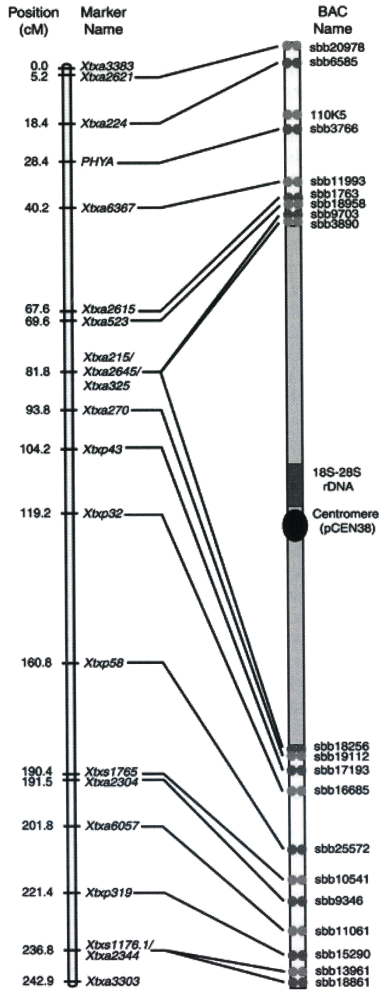
# Construction de la carte

## Cartographier

Pour chaque groupe de liaison (partie de chromosome), déterminer l'ordre des marqueurs et les distances (taux de recombinaisons) qui séparent deux marqueurs adjacents

## Carte saturée

autant de groupes que de chromosomes, tous les marqueurs de la carte sont liés à un groupe.



# Trouver une bonne carte

## Un problème combinatoire

Pour $n$ marqueurs, il y a $n!/2$ ordres de marqueurs définissant des cartes différentes. $\frac{10!}{2} = 1.810^6$.

- Impossible d'énumérer les ordres.
- Problème d'optimisation difficile (même dans ses versions les plus simples).

# Logiciels de cartographie

- végétaux: MapMaker, CarthaGene, JoinMap
- animaux: CRIMAP,
- homme: MapMaker
- hybrides irradiés: RHMAP, RHO, CarthaGene

Voir http://linkage.rockefeller.edu/soft/