# Introduction à l'épidémiologie génétique

**Hervé Perdry – Février 2019**

herve.perdry@u-psud.fr

# Plan

**Génétique élémentaire**

**Maladies monogéniques**

**Traits quantitatifs et maladies complexes**

# Génétique élémentaire

**Lois de Mendel**

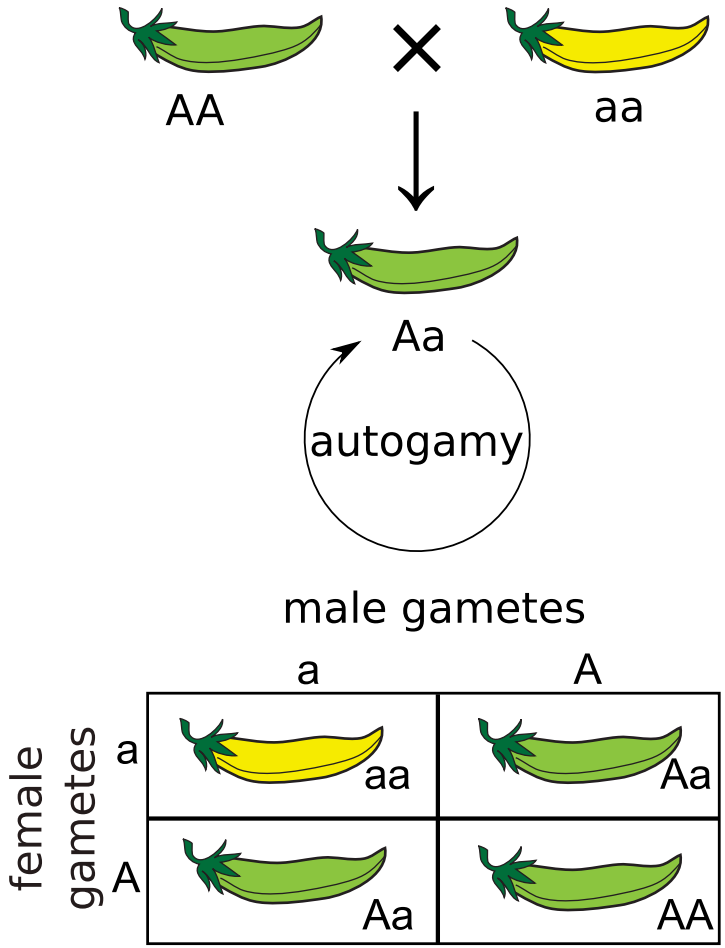
**Chromosomes**

**Recombinaison et distance génétique**

**Polymorphismes génétiques**

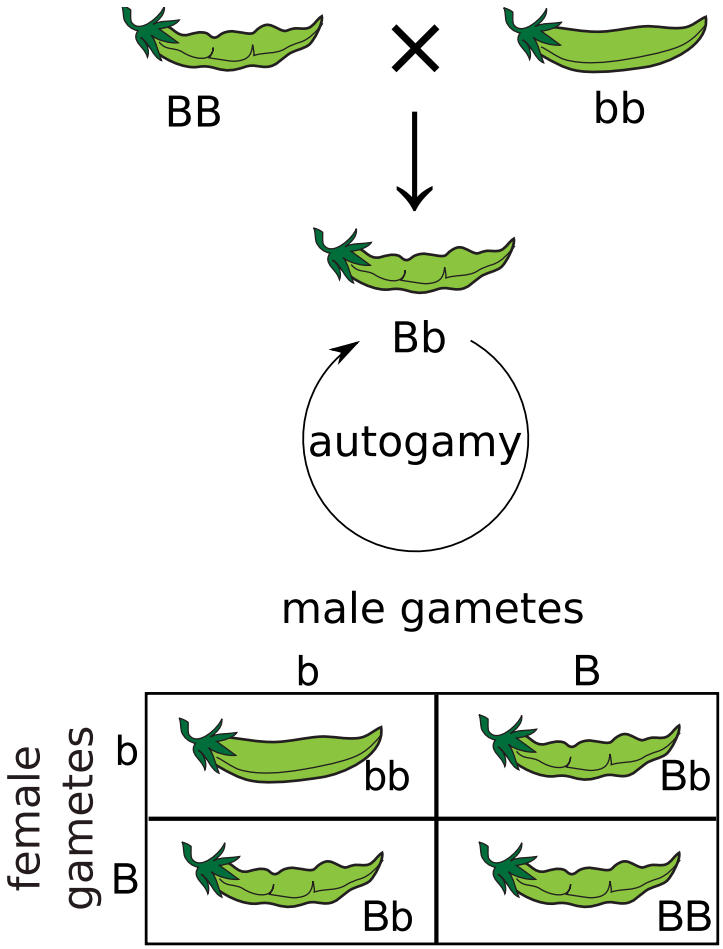
# Lois de Mendel

## Première loi : ségrégation des caractères



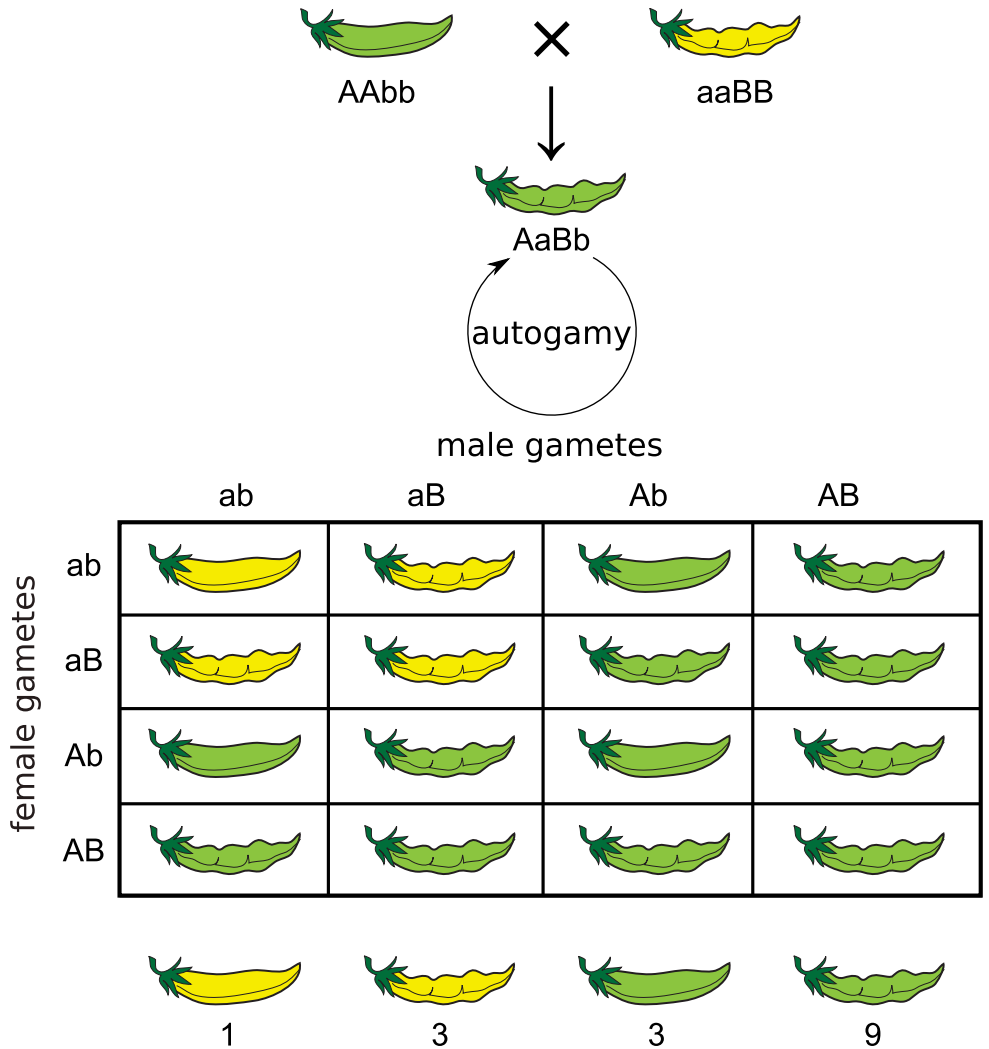
# Lois de Mendel

## Première loi : ségrégation des caractères



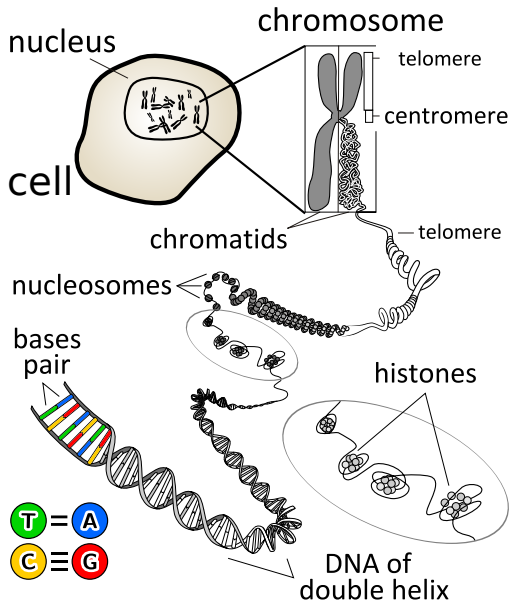
# Lois de Mendel

## Deuxième loi : ségrégation indépendante



# Chromosomes

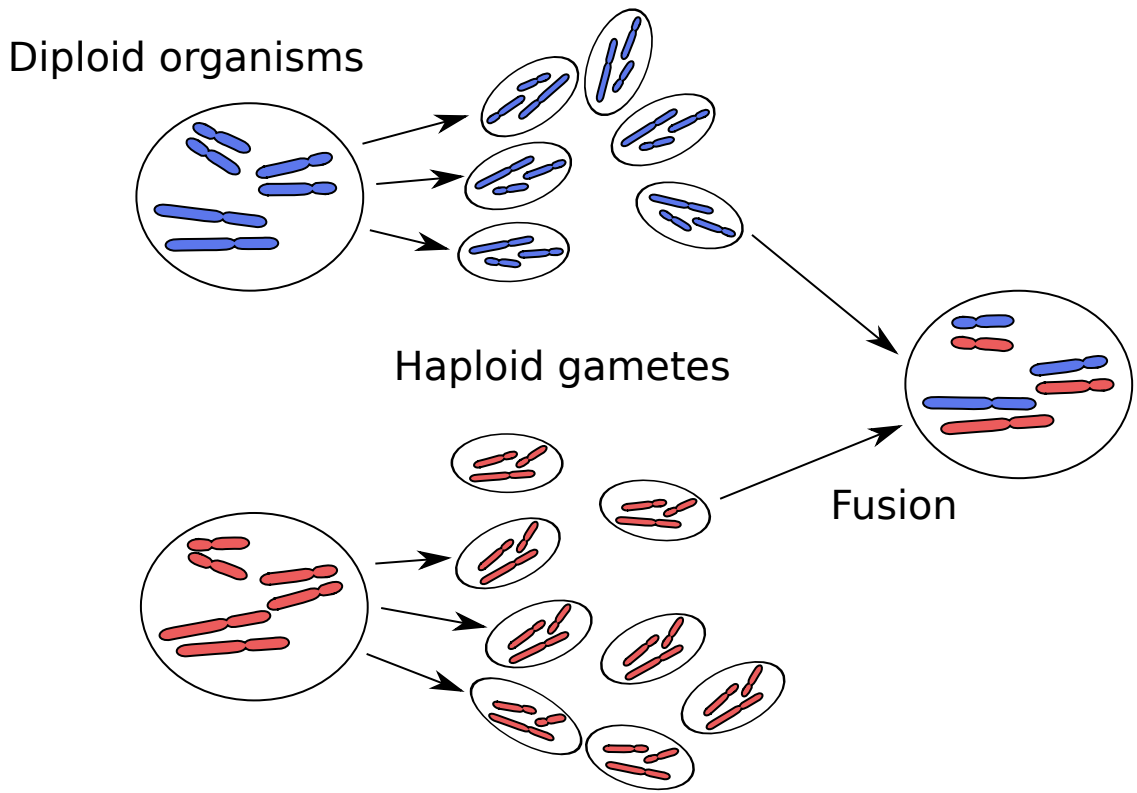
Les chromosomes se trouvent dans le noyau de la cellule ; ils contiennent deux longues chaînes d'ADN fait de quatre types de nucléotides (ACGT). La séquence de ces quatre nucléotiques forme l'information génétique.




Les humains (entre autres...) sont diploïdes, c'est-à-dire que nous avons deux copies de chaque chromosome, *à l'exception des chromosomes sexuels X/Y*. Longueur totale du génome : 3.1 Gb, sur les 22 paires d'autosomes et la paire X/Y.

# Reproduction et gamètes

Chaque parent diploïde contribue à la moitié du génome de l'enfant à travers des gamètes haploïdes.



# Coségrégration ou indépendence ?

## Rétrocroisement (backcross)

Quelques expériences de Thomas Hunt Morgan.

- On croise une drosophile sauvage (génotype AA) avec un individu mutant présentant un trait récessif (génotype aa)
  ☞ individu hybride (génotype Aa, phénotype dominant)
- on croise l'hybride avec son parent mutant (backcross) – ou avec un invidividu de génotype aa
- Les individus résultants sont Aa (phénotype dominant) ou aa (récessif)
  ☞ on peut vérifier les proportions théoriques (première loi, 50 % de chaque)

# Coségrégration ou indépendence ?

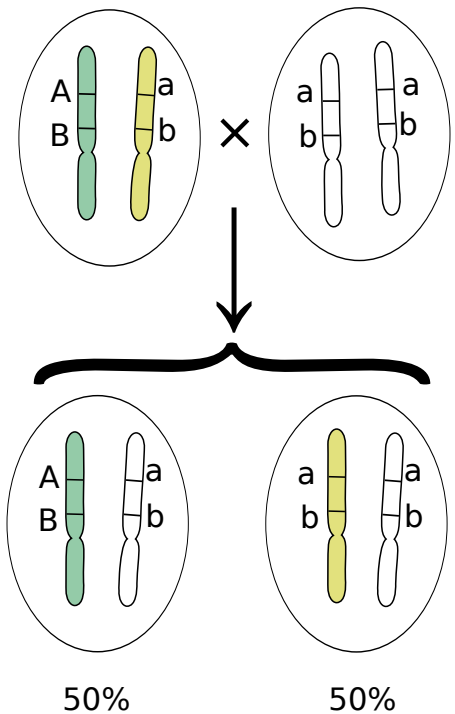
## Double rétrocroisement (double backcross)

- On croise une drosophile sauvage (génotype AA, BB) avec un individu mutant présentant deux traits récessifs (génotype aa, bb)
  - ☞ individu di-hybride (génotype Aa, Bb) ;
- Backcross avec le parent récessif
- Les individus résultants sont Aa ou aa au premier locus, Bb ou bb au second
  - ☞ quatre types possibles (Aa Bb, aa Bb, aa bb ou Aa bb)
    - selon la seconde loi, 25% de chaque type
    - mais est-ce possible si les deux locus sont sur le même chromosome ?

# Coségrégration ou indépendence ?

## Théorie (naïve) de l'hérédité chromosomique

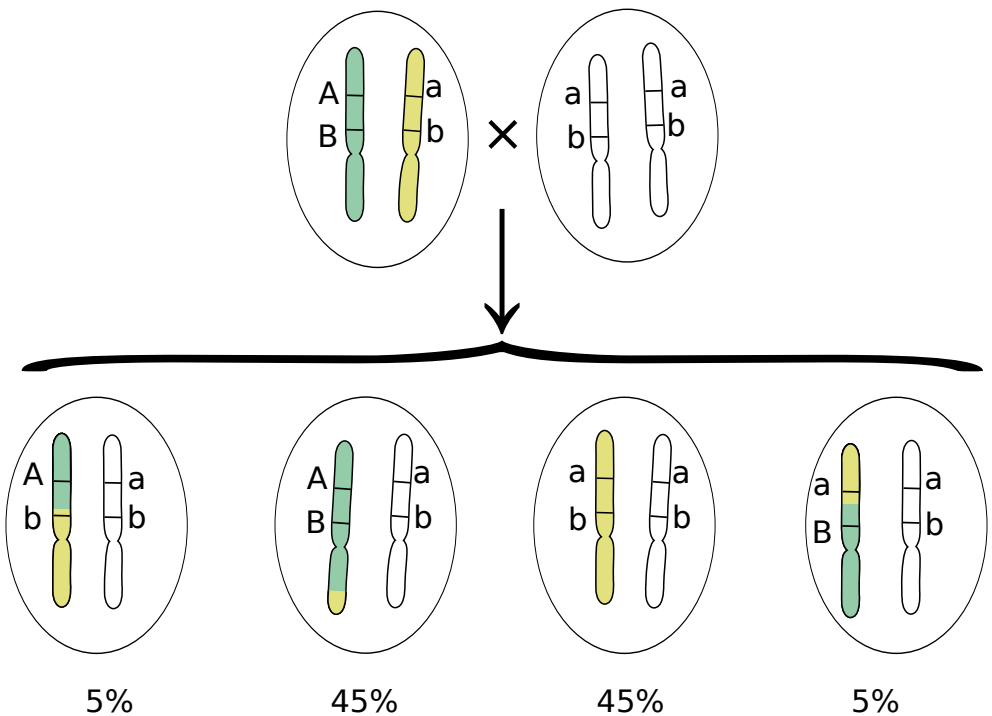
Si deux gènes sont sur le même chromosome, les allèles devraient coségréger :



# Coségrégration ou indépendence ?

## Recombinants

En réalité, on observe des individus recombinants (ici, $\theta = 10\%$ de recombinants).

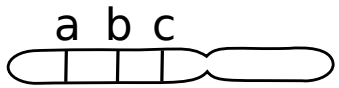


# Distance génétique

- le taux de recombinaison $\theta$ peut être utilisé pour mesurer la « distance génétique » entre deux gènes / deux locus
- si deux locus ne sont pas génétiquement liés (par ex. pas sur le même chromosome), $\theta = 50\%$ (et la seconde loi de Mendel est valide)
- si il y a deux événements de recombinaison entre deux locus, du point de vue du phénotype observé c'est comme si il n'y avait pas eu de recombinaison
  - ☞ $\theta$ est la probabilité d'un *nombre impair* d'enjambements
    On a $0 \leq \theta \leq 0.5$ (indépendance des enjambements)
- en utilisant des tri-hybrides, on peut ordonner trois locus sur un chromosome, et finallement construire une carte du génome

# Distance génétique

- Unité de distance génétique : Morgan (M) ou centiMorgan (cM).
  - Un Morgan = en moyenne, un enjambement par génération
  - Un cM = en moyenne, un enjambement toutes les 100 générations
- Sur le génome humain, 1cM $\simeq$ 1 Mb
- Pour les petites distances, $\simeq$ taux de recombinaison $\theta$
- Ne fonctionne pas sur les grandes distance ; ou encore, les taux de recombinaison ne sont pas additifs :




$$\theta_{ac} \neq \theta_{ab} + \theta_{bc}$$

# Distance génétique

## Carte de Haldane

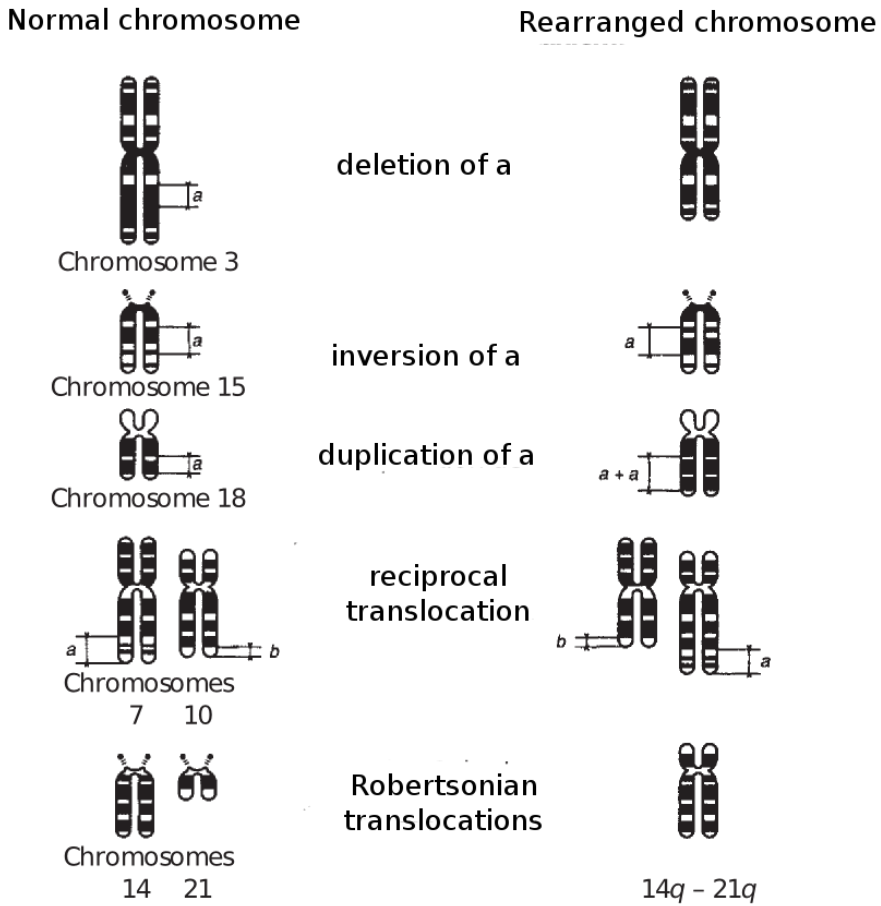
- ☞ On peut relier le taux de recombinaison à la distance génétique par la transformation suivante :

$$d(a,b) = -\frac{1}{2}\log(1-2\theta_{ab})$$

  - Si $\theta_{ab}$ est petit, $d(a,b) \simeq \theta_{ab}$
  - Ceci est lié au fait que le taux de recombinaison est la probabilité d'avoir un nombre *impair* d'enjambements entre les deux locus considérés (deux enjambements : pas de recombinaison) et à l'indépendance (supposée) entre les enjambements

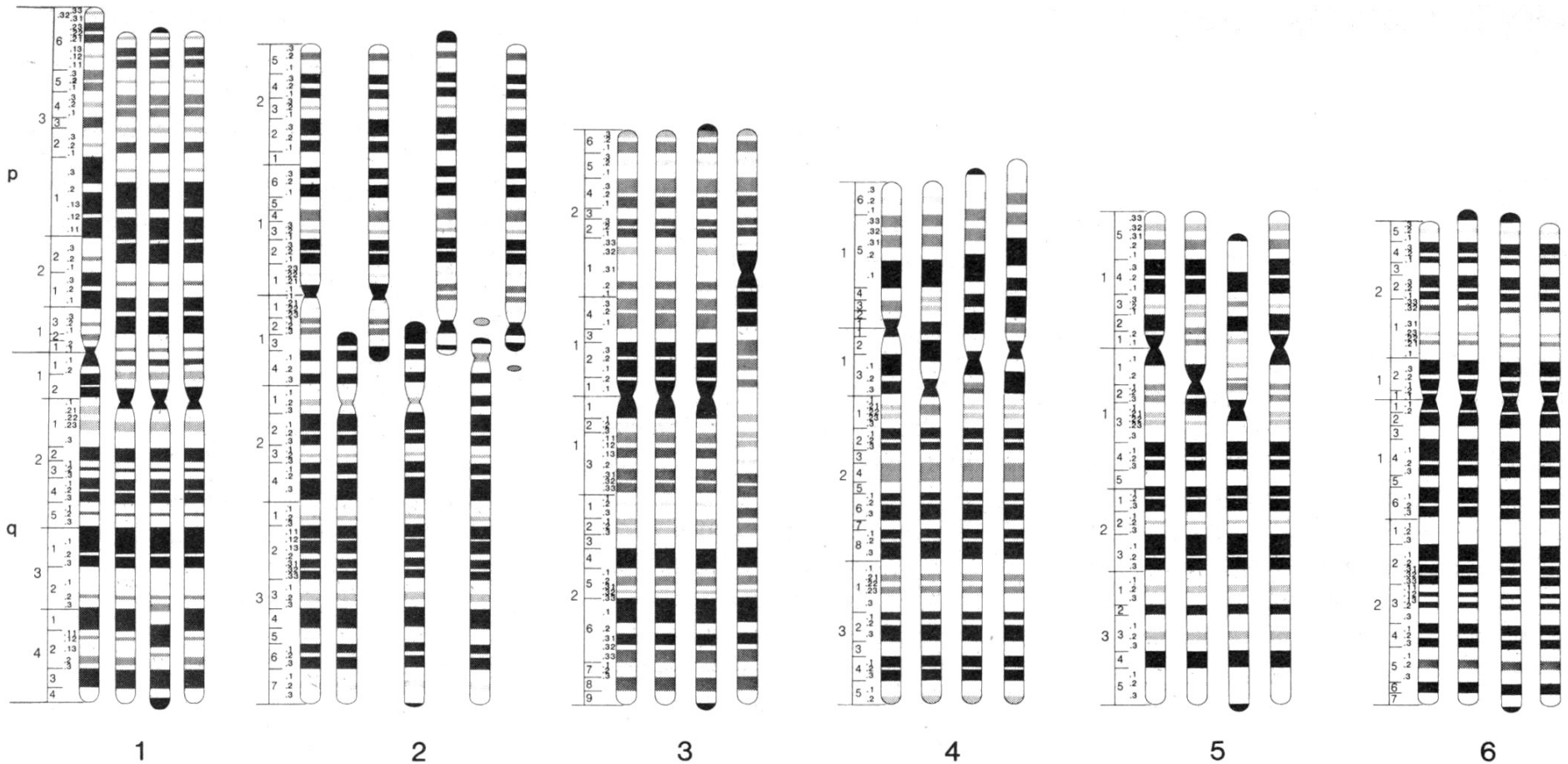
# Polymorphisme génétique

Le polymorphisme peut affecter des parties de chromosomes entières.



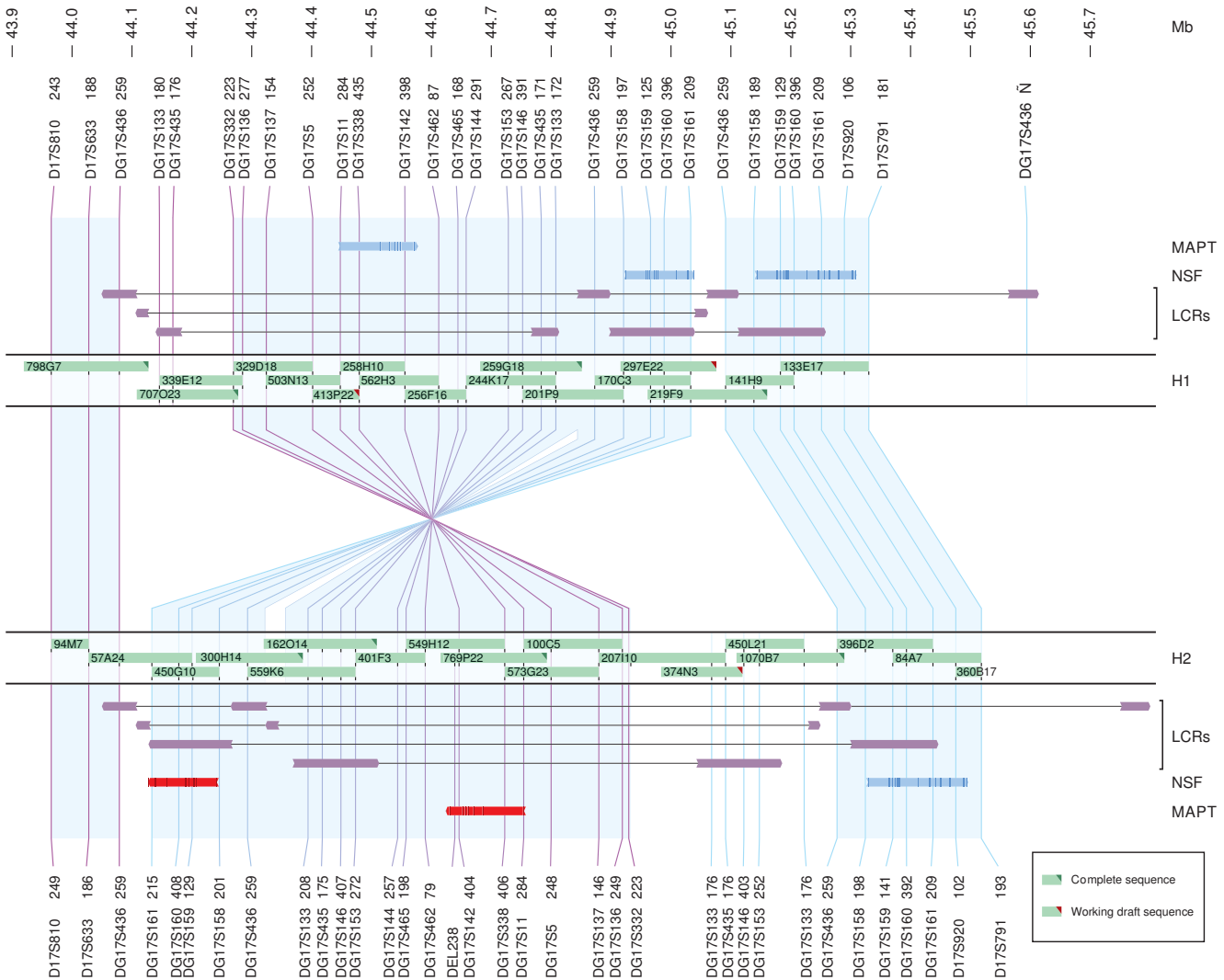
# Polymorphisme génétique

## Caryotypes humain/chimpanzé/gorille/orang-outan



# Polymorphisme génétique

## Inversion (A common inversion under selection in Europeans, Stefanson 2005)




A 900-kb inversion polymorphism was detected on one of the RP11 chromosomes at the 17q21.31 locus. The breakpoints reside in 100-kb palindrome low-copy repeats (LCRs), 900kb apart.

# Polymorphisme génétique

## Au niveau protéique

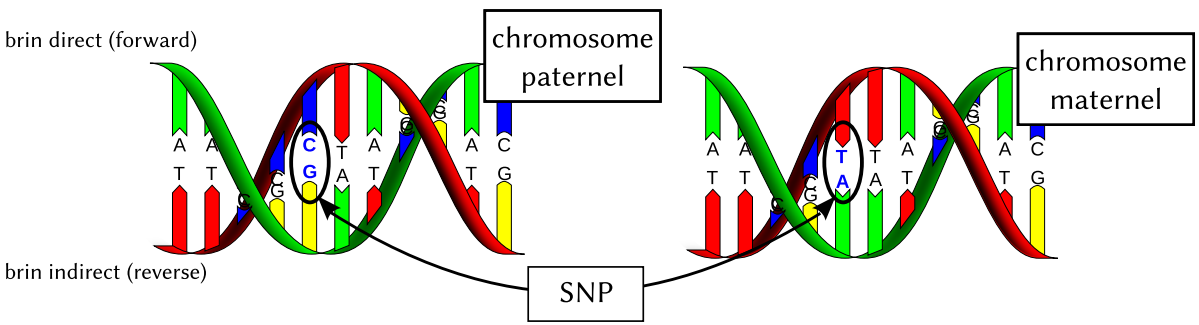
Le polymorphisme génétique peut être réflété au niveau des protéines, qui seront présentes ou non, sous diverses isoformes. Ce polymorphisme peut être révélé directement par diverses techniques (antigènes, électrophorèse)

- groupe sanguin ABO
  Gène ABO. Les *allèles* de ce gène sont notés $I^A$, $I^B$ et $i$.
- sérotype HLA
- etc.

# Polymorphisme génétique

## Au niveau de l'ADN

- Microsatellites, Variable Number Tandem Repeats
  Courts motifs (1 to 4 bp) répétés.
  Cas particuliers : répétition de triplets CAG (Huntington, dans région codante) and CGG (X-fragile).
- Polymorphisme affectant seulement une paire de base : microdéletion et insertion (frameshift), Single Nucleotide Polymorphism (SNP)
  Ce sont les polymorphismes les plus fréquents (approx. 15 millions connus).



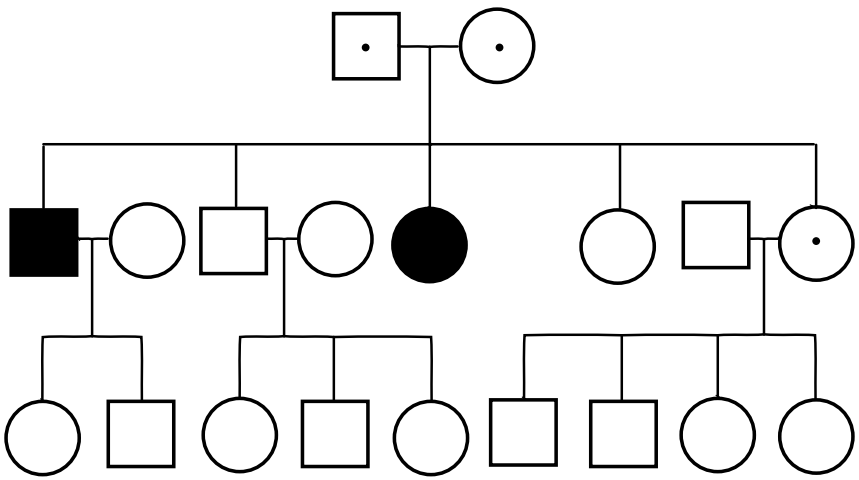

Si le brin indirect (reverse strand) est choisi comme référence, le génotype pour ce SNP peut être CC, CT or TT (GG, GA or AA sur le brin direct).
Récodé souvent en 0, 1, 2 dans les fichiers de données génétiques.

# Maladies monogéniques

- Modèles
- Analyse de ségrégation
- Analyse de liaison
- Localisation par autozygotie
- Données de séquences : recherche de mutations délétères rares / de novo

Maladies récessives.

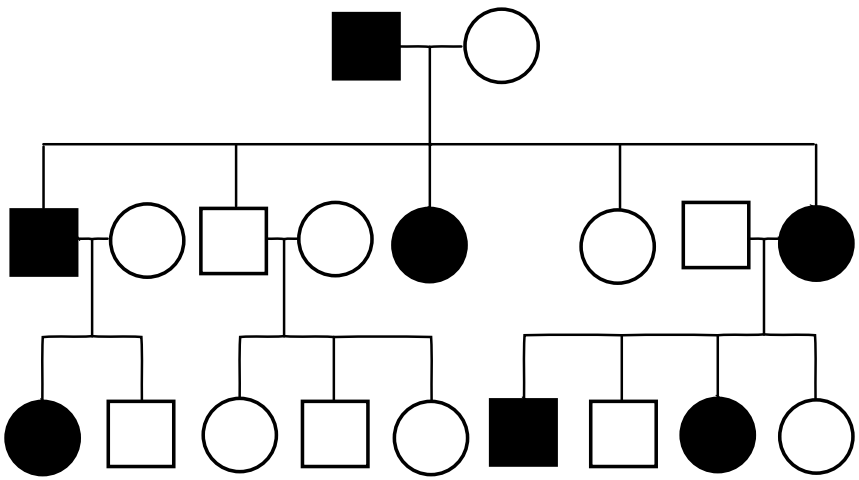
Dans les fratries 25% d'atteints si les deux parents sont porteurs.



# Modèles de maladie

## Maladies mendéliennes

Maladies dominantes.
Un parent atteint transmet la maladie à 50 % de ses enfants.



$\oplus$ maladies liées à l'X ou à l'Y, maladies mitochondriales.

# Modèles de maladie

## Modèles plus complexes

- Maladie monogénique avec phénocopies
  - Individus qui n'ont pas la mutation causale mais présentent des symptômes similaires
- Maladie monogénique avec pénétrance incomplète
  - Individus qui ont la mutation causale mais ne présentent pas de symptômes

☞ Difficulté à reconnaître le mode de transmission

- Hétérogénéité : plusieurs gènes causent la « même » maladie
  - Retinitis pigmentosa : plus de 30 gènes ont été identifiés, la plupart dominants ; un est récessif, un autre récessif lié à l'X.
  - Surdité : formes récessives, dominantes...

☞ Difficuté à obtenir des échantillons homogènes

# Modèles de maladie

## Modèles plus complexes

La pénétrance d'un génotype est la probabilité qu'un individu qui porte ce génotype soit affecté. Pour un locus di-allélique A/a, on pose $f_{AA} = P(aff|AA)$, $f_{Aa} = P(aff|Aa)$, $f_{aa} = P(aff|aa)$.

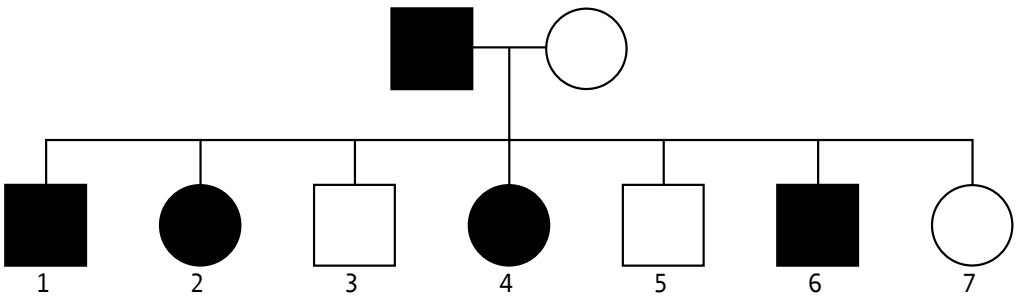
| Modèle | $f_{AA}$ | $f_{Aa}$ | $f_{aa}$ |
|---|---|---|---|
| Récessif | 0 | 0 | 1 |
| Dominant | 0 | 1 | 1 |
| Recessif avec phénocopies | 0.001 | 0.003 | 1 |
| Recessive avec pénétrance incomplète | 0 | 0 | 0.85 |
| Effet majeur | 0.001 | 0.1 | 0.3 |
| (modèle polygénique) Effet faible | 0.001 | 0.0015 | 0.003 |

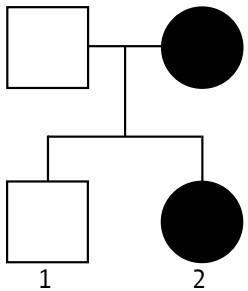
Les risques relatifs sont souvent utiles

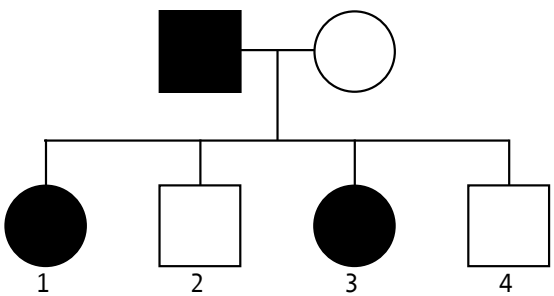
| Model | $RR_{AA}$ | $RR_{Aa}$ | $RR_{aa}$ |
|---|---|---|---|
| Effet majeur | 1 | 100 | 300 |
| Effet faible | 1 | 1.5 | 3 |

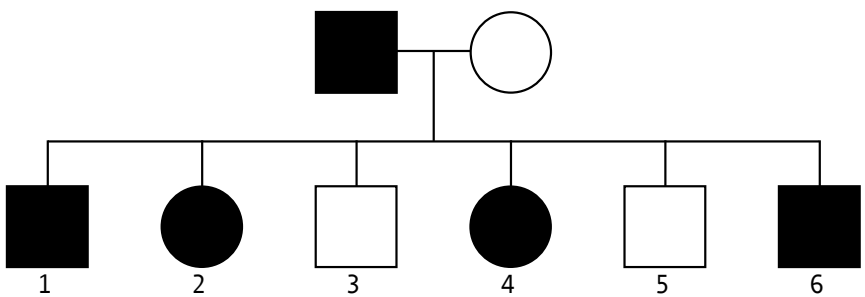
# Analyse de ségrégation

Devinette : que pouvez-vous déduire de ces généalogies ?



# Analyse de ségrégation

## Réponse

- La maladie n'est pas liée à l'X ou à l'Y
- Transmise à approximativement un enfant sur deux
- ☞ Transmission dominante (autosomale)

C'est exactement ce que fait l'analyse de ségrégation : inférer un modèle de maladie à partir de généalogies. Ceci permet notamment de prouver l'existence d'une composante génétique dans le trait étudié.

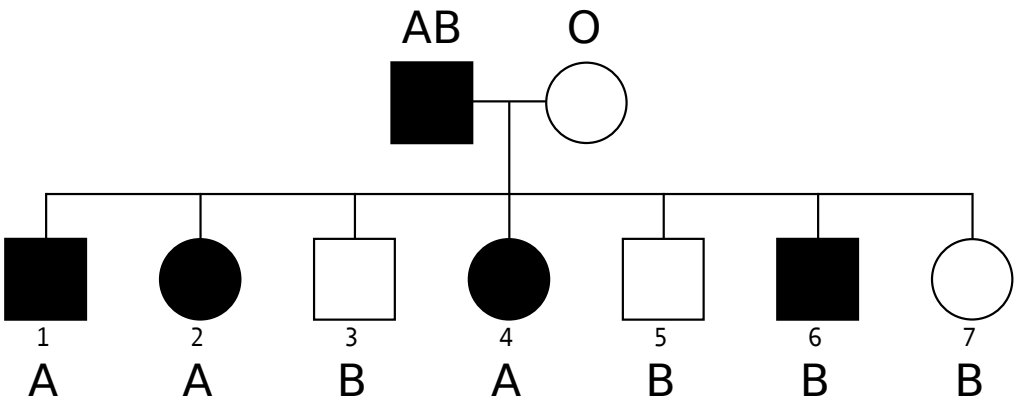
Modèles

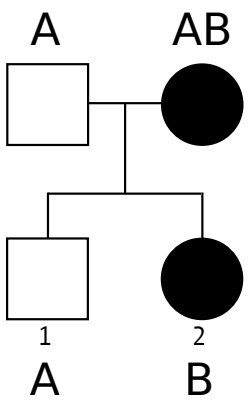
- Mendélien
- Monogénique avec trois pénétrances $f_{AA}$, $f_{Aa}$, $f_{aa}$
- Possibilité de modéliser l'environnement familial, une composante polygénique....

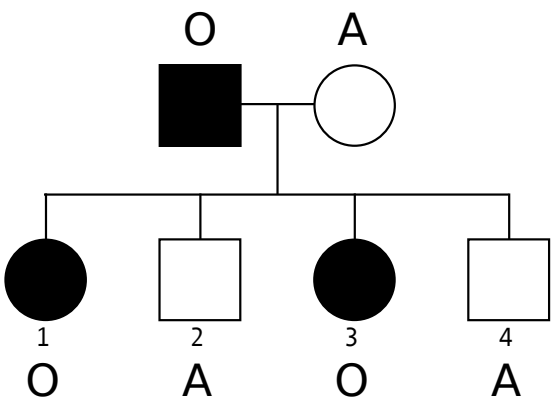
Méthodes statistiques de choix de modèle / estimation des paramètres...

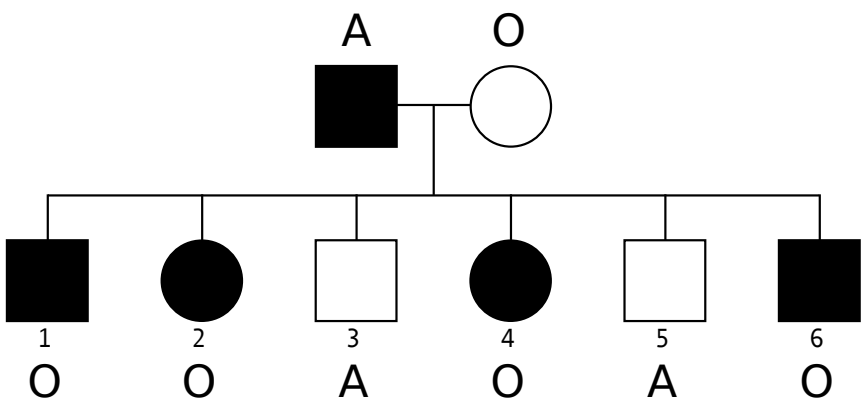
# Analyse de liaison

Devinette : que pouvez-vous déduire de ces généalogies ?



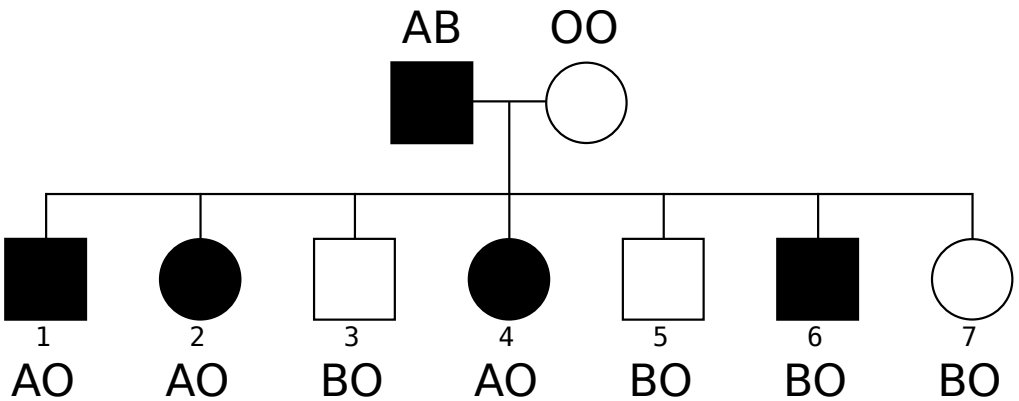



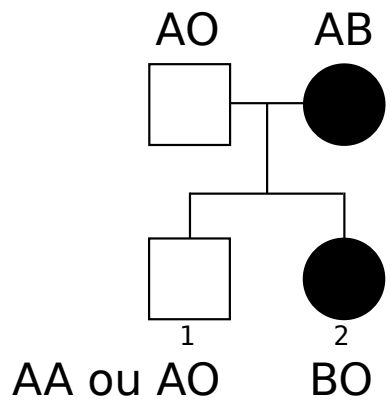



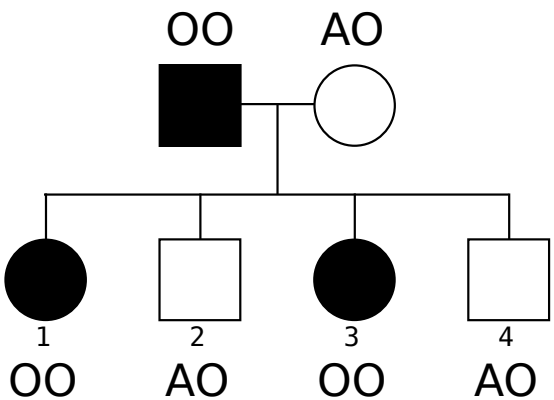



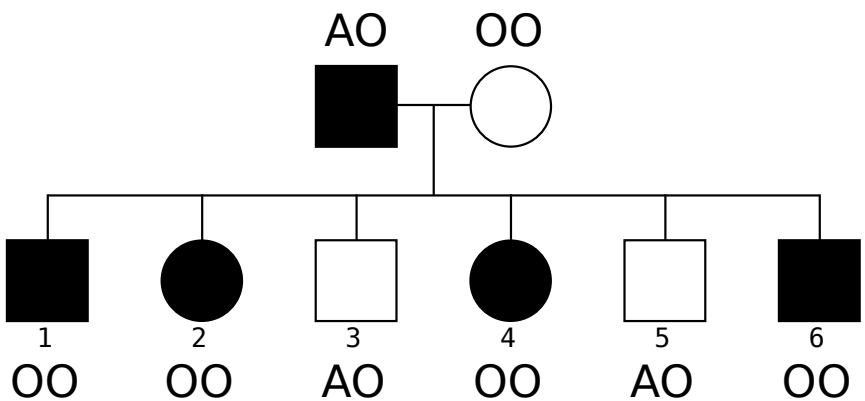
# Analyse de liaison

Devinette : que pouvez-vous déduire de ces généalogies ?



# Analyse de liaison

## Réponse

- Nous avons déjà déterminé que la transmission est dominante
- La maladie est co-transmise avec un allèle du locus ABO
  - ☞ Liaison génétique avec le locus ABO
- Grosso modo un recombinant parmi 15 méioses informatives
  - ☞ Le taux de recombinaison entre le locus morbide et le locus ABO peut être estimé par $\theta = 1/15$ (donc 7 cM entre le locus ABO et le locus maladie)
- Pour répondre à cette devinette, nous avons
  - inféré le génotype de chaque individu au locus maladie, en utilisant le modèle de la maladie (maladie dominante)
  - inféré si il y avait eu recombinaison entre le locus maladie et le marqueur génétique (ABO)

C'est l'analyse de liaison classique, dite « paramétrique ».

# Analyse de liaison

Le principe de l'analyse de liaison est d'estimer la distance génétique (centiMorgans) ou le taux de recombinaison entre le locus maladie hypothétique et des « marqueurs » de position connue sur le génome.

- Nécessité d'un modèle pour la maladie (dominante, récessive, etc)
- Nécessité d'une carte du génome
- La résolution est limitée par le nombre de méïoses observées : on ne peut pas séparer un marqueur du locus si on n'observe pas de recombinaison

# Analyse de liaison sur tout le génome

- En utilisant des marqueurs qui couvrent le génome (environ 1000 microsatellites), on peut identifier la région où se trouve le locus maladie.
- Plus les marqueurs sont polymorphes, mieux ça marche : on utilisait classiquement des micro-satellites
- Des méthodes ont été dévelopées pour utiliser des SNPs (données hégémoniques aujourd'hui)

# Après l'analyse de liaison

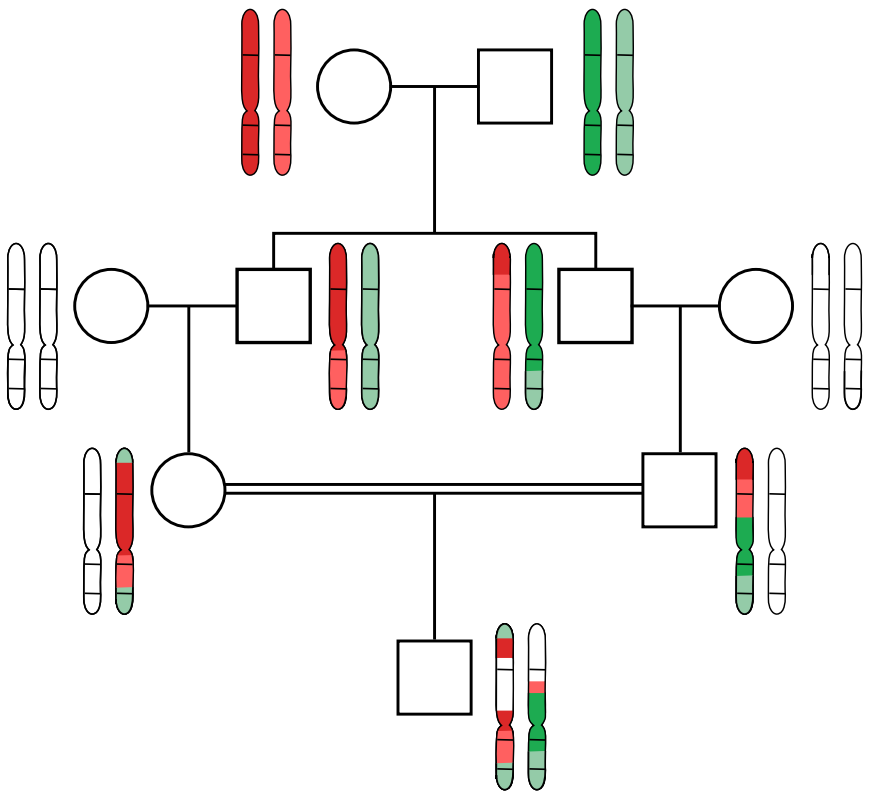
Une fois identifiée la région du génome où se trouve le gène responsable de la maladie, reste à l'identifier avec précision.

- Analyse d'association / séquençage des gènes chez les cas
- Fonction des gènes
- Modèles animaux
- Modèles in vitro
- etc

# Localisation par autozygotie

## Consanguinité et maladies récessives

Des individus consanguins peuvent être autozygotes (ou Homozygous by Descent : HBD) sur tout un segment chromosomique.



En un locus donné, l'enfant a une probabilité $f = \frac{1}{16}$ d'être HBD.
$f$ est appelé coefficient de consanguinité de l'individu.

# Localisation par autozygotie

## Consanguinité et maladies récessives

Probabilité d'un individu avec consanguinité $f$ d'avoir un génotype AA :

- si il est HBD au locus considéré (probability $f$), il est AA avec probabilité $p$
- s'il n'est pas HBD (probabilité $1-f$), il est AA avec probabilité $p^2$

☞ $P(AA) = (1-f)p^2 + fp$

De même

- $P(Aa) = (1-f)2pq$
- $P(aa) = (1-f)q^2 + fq$

Mêmes formules pour la population globale en prenant $f =$ consanguinité moyenne.

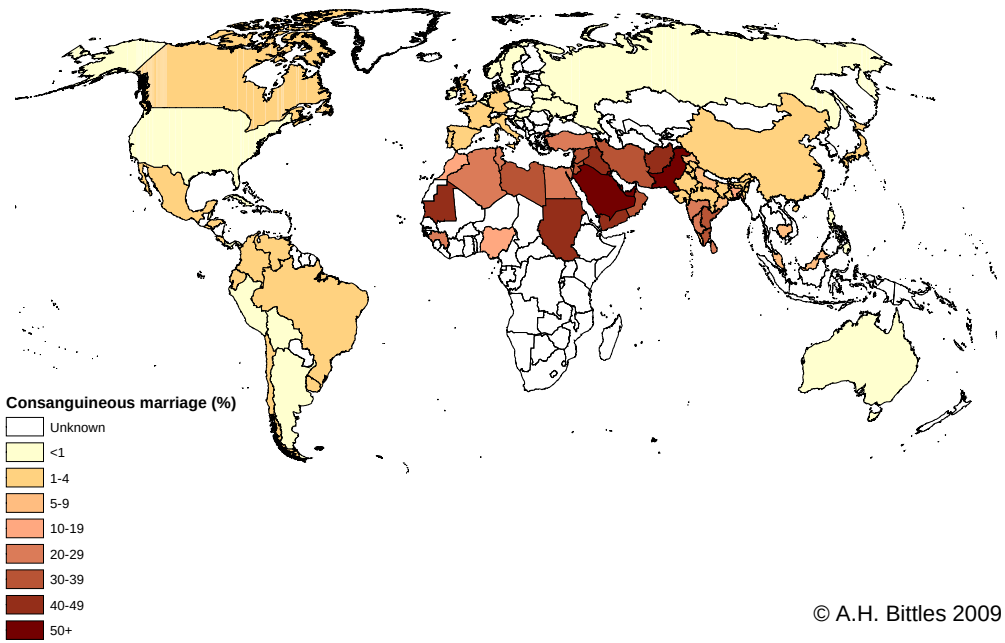
# Localisation par autozygotie

## Consanguinité et maladies récessives

La fréquence des maladies récessives (rares) est plus grande ches les individus consanguins :

$$(1-f)q^2 + fq \gg q^2 \text{ for small } q$$

- Garrod (1902) a remarqué que ses sujets alcaptonuriques étaient enfants de cousins germains.



# Localisation par autozygotie

## Mise en œuvre

- Chez les patients consanguins, le gène responsable d'une maladie récessive doit se trouver dans une région autozygote
- Il y a homozygotie en tous les polymorphismes d'une région autozygote
- ☞ On peut identifier les régions autozygotes comme des régions homozygotes « anormalement longues »...
- Ensuite, comme dans le cas de l'analyse de liaison, il faut rechercher quel gène de la région est responsable
- De cette façon on peut identifier des gènes responsables de maladies récessives rares, parfois avec un seul patient
- Utilisation de microsatellites, SNPs, données de séquence...

# Recherche de mutations délétères rares ou de novo

- Grâce aux technologies de séquençage à haut-débit, on peut séquencer la totalité de l'exome pour un prix « abordable »(et bientôt, la totalité du génome)

☞ Méthodes bio-informatiques pour prédire si un variant est délétère (e.g. modifie la structure de la protéine)

☞ Recherche de variants délétères partagés par les patients (ou présents dans les mêmes gènes, etc)

- En séquençant un individu et ses deux parents on peut rechercher des mutations de novo sur tout l'exome...
- Attention, mutations de novo non-sens ou faux-sens dans l'exome : selon les études, entre 0,5 et 1 en moyenne par individu sain !

# Traits quantitatifs et maladies complexes

- Modèles
- Héritabilité et récurrence familiale
- Analyse d'association

# Modèle polygénique

## Traits quantitatifs

- Comment modéliser des traits quantitatifs ? (niveau de cholestérol, taille, etc.)
  - Les premières modélisations de la corrélation entre apparentés prédatent la redécouverte des lois de Mendel en 1900
  - La multiplicité de valeurs de ces traits rend à première vue le modèle mendélien inapproprié

- ☞ Modèle polygénique (Fisher 1918)
  - Un grand nombre de gènes intervient
  - Chacun a un effet faible
  - Les effets alléliques sont additifs (co-dominance)
  - Pas d'épistasis, c.-à-d. pas d'interaction entre gènes
  - Les gènes ne sont pas liés (ségrégation indépendante)
  - Ni interaction ni corrélation gène-environnement

# Modèle polygénique

## Traits quantitatifs

Dans ce modèle on décompose le phénotype $P$ en une composante génétique $G$ et une composante environnementale $E$ :

$$P = G + E$$

- $G$ est l'effet total du génome, c'est la somme d'un grand nombre de petits effets indépendants
  - ☞ Normal, variance $\sigma_G^2$
- $E$ est l'effet environnemental (ainsi qu'une composante aléatoire)
  également supposé normal, variance $\sigma_E^2$

En supposant que la composante génétique $G$ et la composante environnementale $E$ sont indépendantes, on a

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

L'héritabilité est définie par $h^2 = \sigma_G^2/\sigma_P^2$. C'est la proportion de la variance du trait qui est due à la composante génétique.

# Estimer l'héritabilité des traits quantitatifs

## Études de jumeaux

Cette estimation a le même but que l'analyse de ségrégation dans le cas monogénique : prouver l'existence d'une composante génétique dans le trait étudié.

- Utiliser des données familiales pour estimer $h^2$
- Le cas le plus simple est celui de jumeaux monozygotes :
  - $P_1 = G + E_1$
  - $P_2 = G + E_2$
  - ☞ $\text{cov}(P_1, P_2) = \sigma^2_G$

☞ en utilisant un échantillon de jumeaux MZ, on peut estimer l'héritabilité $h^2 = \sigma^2_G/\sigma^2_P$ par $\text{cov}(P_1, P_2)/\text{var}(P) = \text{cor}(P_1, P_2)$.

Oui ?

# Héritabilité des traits quantitatifs

## Études de jumeaux

- Les jumeaux monozygotes partagent une partie de leur environnement !
  - $P_1^{MZ} = G + E_S + E_1$
  - $P_2^{MZ} = G + E_S + E_2$
  - ☞ $\text{cov}(P_1^{MZ}, P_2^{MZ}) = \sigma^2_G + \sigma^2_{E_S}$
- En supposant que les jumeaux dizygotes partagent leur environnement dans les même proportions :
  - $P_1^{DZ} = G_1 + E_S + E_1$
  - $P_2^{DZ} = G_2 + E_S + E_2$
  - ☞ $\text{cov}(P_1^{DZ}, P_2^{DZ}) = \text{cov}(G_1, G_2) + \sigma^2_{E_S} = \frac{1}{2}\sigma^2_G + \sigma^2_{E_S}$

☞ Estimation de $\sigma^2_G$ par $2(\text{cov}(P_1^{MZ}, P_2^{MZ}) - \text{cov}(P_1^{DZ}, P_2^{DZ}))$

On en tire une estimation de $h^2$ par

$$2\big[\text{cor}(P_1^{MZ}, P_2^{MZ}) - \text{cor}(P_1^{DZ}, P_2^{DZ})\big]$$

c'est-à-dire deux fois la différence des corrélations entre jumeaux MZ et DZ.

☞ Une plus grande corrélation entre jumeaux MZ qu'entre jumeaux DZ montre que le trait a une composante génétique.

Cependant, la supposition que les jumeaux DZ partagent leur environnement dans la même proportion que les jumeaux MZ est discutable.

Si les jumeaux MZ partagent plus d'environnement que les DZ, on surestime $h^2$.

# Héritabilité des traits quantitatifs

- D'autres corrélations familiales peuvent être utilisées également

  ☞ nécessité de modéliser la façon dont l'environnement est partagé

- Récemment, des méthodes utilisant des données génomiques d'individus non-apparentés ont été proposées (Visscher).

  Elle n'éliminent pas nécessairement tous les biais !

# Modèle polygénique

## Critiques du modèle polygénique et de l'héritabilité

Un concept flou, défini dans un modèle simpliste, difficile à mesurer et à interpréter

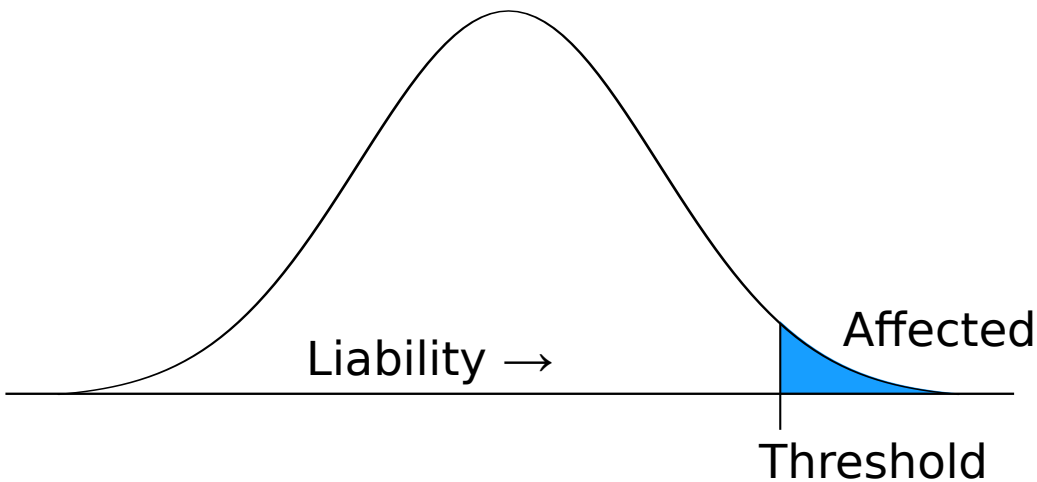
- Existence de facteurs environnementaux à forts effets (infections, expositions) : remet en cause le modèle gaussien pour $E$
- Interactions $G \times E$ : certains gènes n'ont un rôle que dans un environnement donné (voire des rôles différents dans des environnements différents)
- Corrélations gène-environnement (à toutes les échelles)
- Valeur propre à une population, pas « transférable »
- Si $\sigma_E^2$ devient petit, $h^2$ devient grand. Si l'environnement varie peu, l'héritabilité est élevé (cas des animaux d'élevage).
- N'est bon qu'à une chose : prédire le succès de la sélection artificielle...

# Modèle polygénique

## Maladies complexes

Certaines maladies n'ont pas un mode de transmission mendélien clair, mais sont pourtant réputées frapper plus certaines familles que d'autres : maladies cardiovasculaire, cancers, maladies auto-immunes (diabète de type 1, sclérose en plaques, maladie cœliaque), maladies psychiatriques...

Une façon possible d'étendre le modèle polygénique au cas des maladies complexes est le modèle de la « liabilité » (liability), qui suppose l'existence d'une quantité connue « cachée » (la liabilité), avec un seuil au-delà duquel la maladie se déclare.



# Modèle polygénique

## Maladies complexes

Le modèle de la liabilité est (presque) équivalent au

- modèle de la régression logistique :

$$\text{logit } P(Aff) = \alpha + \beta_1 SNP_1 + \cdots + \beta_n SNP_n$$

où $\text{logit } p = \log \frac{p}{1-p}$.

Pour relier les deux, on prend comme liabilité
$\alpha + \beta_1 SNP_1 + \cdots + \beta_n SNP_n + \varepsilon$ où $\varepsilon$ est un terme d'erreur (variable aléatoire)

L'avantage du modèle logistique est que les $\beta_i$ s'interprètent naturellement comme des logarithmes d'odds ratios (risques relatifs).

# Modèle polygénique

## Maladies complexes

Le modèle de la liabilité est (presque) équivalent au

- modèle des risques multiplicatifs, où les risques relatifs aux différents locus se multiplient.

| Gene | Relative risks | | |
|---|---|---|---|
| Gene 1 | AA : 1 | Aa : 1 | aa : 1.3 |
| Gene 2 | BB : 1 | Bb : 1.1 | bb : 1.2 |

| | AA | Aa | aa |
|---|---|---|---|
| BB | 1 | 1 | 1.3 |
| Bb | 1.1 | 1.1 | 1.43 |
| bb | 1.2 | 1.2 | 1.56 |

# Maladies complexes : récurrence familiale et héritabilité

Toujours afin de prouver l'existence d'une composante génétique...

- On recrute des cas au hasard et on considère leurs apparentés
- Un risque plus élevé entre apparentés que dans la population générale est en faveur de l'existence d'une composante génétique
- On peut calculer une héritabilité (modèle de la liabilité)

  Toutes les critiques précédentes s'appliquent, plus d'autres, spécifiques au modèle de la liabilité (hétérogénéité des étiologies, etc).

# Maladies complexes : récurrence familiale et héritabilité

## Exemple de la schizophrénie

| Relation | | Incidence | $\lambda$ |
|---|---|---|---|
| Parents | | 4.36% | 5.45 |
| | valeur corrigée* | 14.12% | 17.65* |
| Germains | | 8.51% | 10.60 |
| Enfants | | 12.31% | 15.40 |
| Oncles, tantes | | 2.01% | 2.50 |
| Demi-germains | | 3.22% | 4.00 |
| Neveux, nièces | | 2.25% | 2.80 |
| Petits-enfants | | 2.81% | 3.50 |
| Cousins | | 2.91% | 3.60 |

$\lambda =$ risque relatif ; calculé pour une incidence en population de 0.8%

* Correction pour le fait qu'une fois la schizophrénie déclarée les patients ont rarement des enfants

Table de Strachan and Read, Human Molecular Genetics, NCBI bookshelf

# Maladies complexes : récurrence familiale et héritabilité

## Exemple des études médicales

| Relation | Prevalence | $\lambda$ |
|---|---|---|
| Germains | 21% | 95 |
| Parents/Enfants | 13% | 60 |

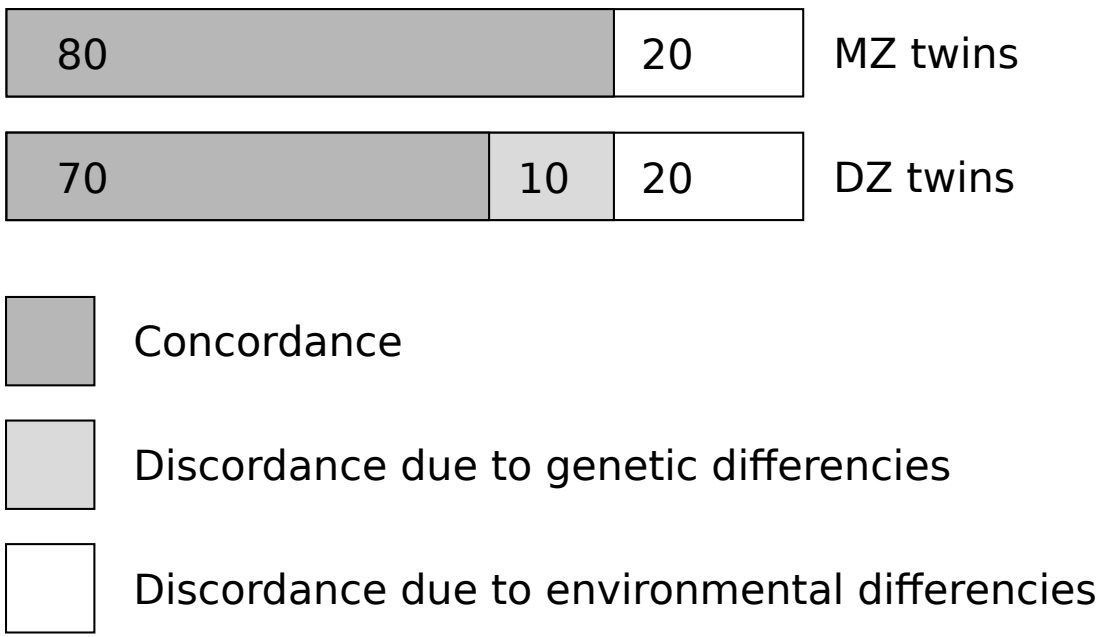
$\lambda =$ risque relatif ; calculé pour une prévalence en population de 0.22%

☞ L'environnement partagé est un facteur de confusion (on peut même avoir des généalogies compatibles avec l'hypothèse d'un trait mendélien)

Huckle and McGuffin, 1990, 1991

# Maladies complexes : études de jumeaux

Même principe qu'avec les traits quantitatifs : les jumeaux MZ et DZ sont supposés partager leur environnement dans les mêmes proportions.

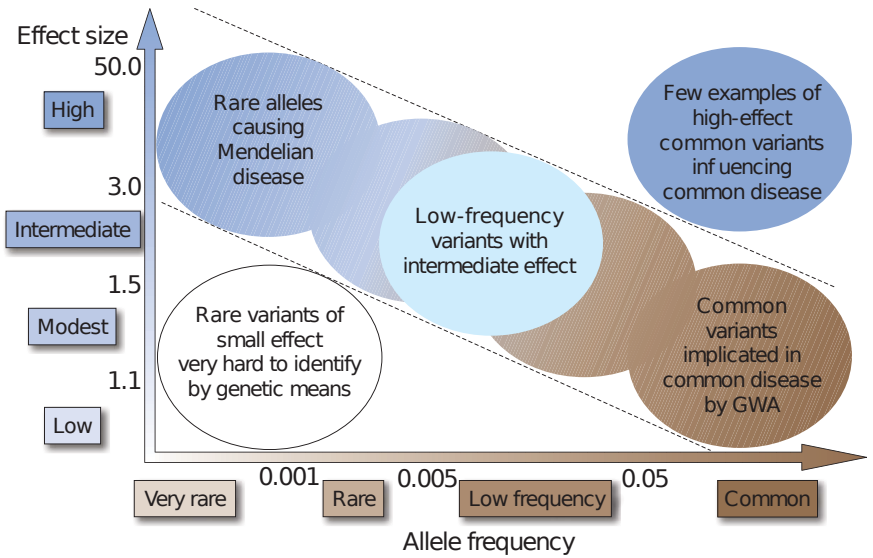



Une concordance plus grande entre jumeaux MZ qu'entre DZ est en faveur de l'existence d'une composante génétique.

# Études de liaison ?

- Les méthodes d'analyse de liaison utilisées pour les maladies monogéniques ne peuvent plus être utilisées
- Des méthodes « disease model free » ont été proposés mais elles n'ont pas donnés des résultats excellents...

Une exception ( ?) : APOE dans la maladie d'Alzheimer




Manolio et al, Nature 2009

# Analyse d'asssociation

## Principes

Les analyses d'association utilisent des données cas/témoins pour déterminer avec précision quels gènes sont impliqués, et pour estimer les risques relatifs.

Exemple : Le gène APOE code la protéine *ApoE*, qui a 4 isoformes, correspondant à 4 alleles du gène : $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$, $\varepsilon_4$. Les deux premiers allèles sont relativement rares dans la population européenne.

| | $\varepsilon_*/\varepsilon_*$ | $\varepsilon_*/\varepsilon_4$ | $\varepsilon_4/\varepsilon_4$ |
|---|---|---|---|
| Cas (Alzheimer) | 37 | 42 | 21 |
| Témoins | 69 | 29 | 2 |

$\varepsilon_* = \varepsilon_2$ ou $\varepsilon_3$.

☞ La répartition des génotypes peut être comparée avec un test du $\chi^2$ (ici la différence est évidemment significative).

# Analyse d'asssociation

- Les études d'associations permettent également l'estimation de risques relatifs associés
- Le plus souvent, études cas/témoin mais il y a aussi des études de cohortes
- Études de gènes candidats (région liée, fonction biologique, etc)
- Dans les années 2000 on est passé aux « analyses d'associations sur tout le génome » (GWAS), avec des tailles d'échantillons de plus en plus importantes.

# Analyse d'asssociation

## Estimation des risques relatifs par odds ratios

| | $\varepsilon_*/\varepsilon_*$ | $\varepsilon_*/\varepsilon_4$ | $\varepsilon_4/\varepsilon_4$ |
|---|---|---|---|
| Cas (Alzheimer) | 37 | 42 | 21 |
| Témoins | 69 | 29 | 2 |

On prend $\varepsilon_*/\varepsilon_*$ comme génotype de référence :

$$\frac{P(Att|\varepsilon_*/\varepsilon_4)}{P(Att|\varepsilon_*/\varepsilon_*)} = \frac{42 \times 69}{37 \times 29} = 2.7$$

$$\frac{P(Att|\varepsilon_4/\varepsilon_4)}{P(Att|\varepsilon_*/\varepsilon_*)} = \frac{21 \times 69}{37 \times 2} = 19.6$$

Si les témoins représentent la population générale les OR estiment un risque relatif

La même analyse peut être réalisée dans le cadre plus général de la régression logistique

# Analyse d'asssociation

## Stratégie de gène candidat

Dans l'approche gène candidat, des gènes sont choisis pour une analyse d'association

- sur la base de leur fonction
- à cause de leur localisation dans une région liée à la maladie
- etc

Si une classification des isoformes d'une protéine encodée par un gène est disponible (exemple d'APOE, groupes sanguins, sérotypes HLA, etc), elle peut être utilisée pour l'analyse.

# Analyse d'asssociation

## Utilisation de SNP

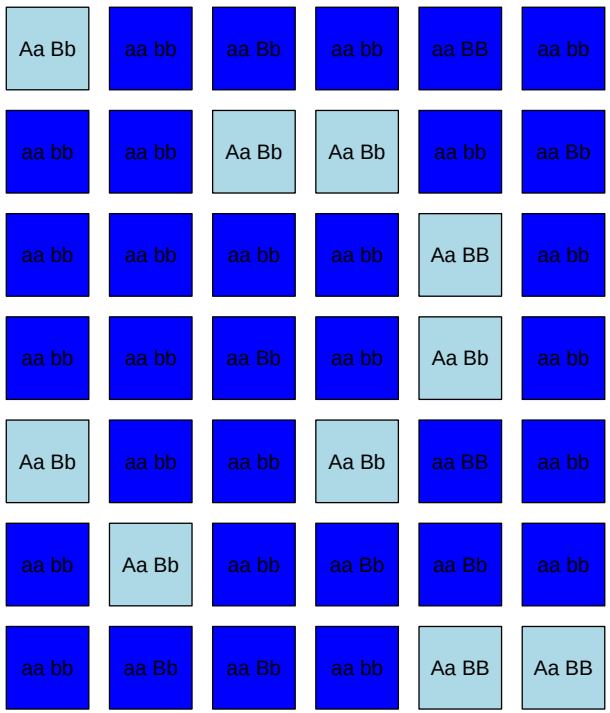
La stratégie la plus générale est d'utiliser des marqueurs (SNP) au sein des gènes ou à leur voisinage.

- Les SNP utilisés ne déterminent pas nécessairement des isoformes distinctes / ne sont pas nécessairement fonctionnels
- Si un SNP est corrélé avec un variant fonctionnel impliqué dans l'analyse, on aura association
  - si un SNP d'allèles A/a est impliqué dans la maladie (allèle a plus fréquent chez les cas)
  - si un SNP voisin d'allèles B/b est corrélé avec le SNP A/a (allèle b plus fréquent quand l'allèle a est présent)
  - on détecte l'effet de A/a à travers l'observation de B/b (allèle b plus fréquent chez les cas)
- Une telle corrélation entre SNP voisins est appelée « déséquilibre de liaison »

# Détection d'une association

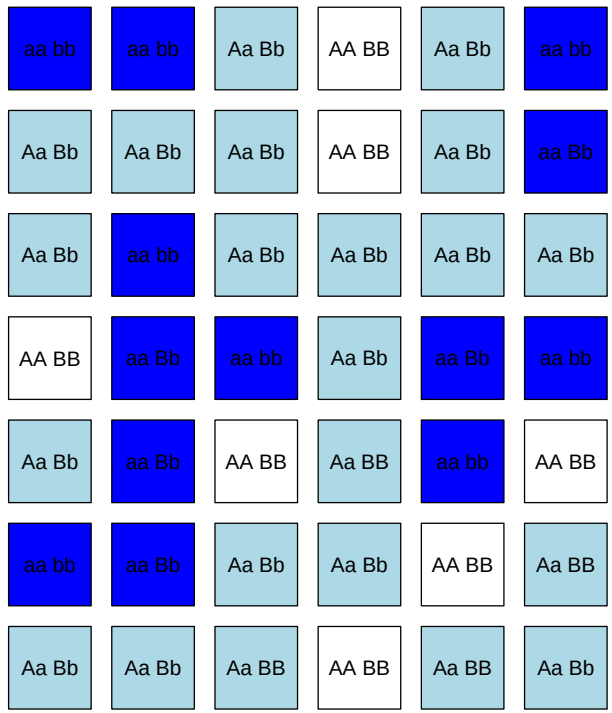
## On observe le SNP impliqué

Cas




| AA | Aa | aa |
|---|---|---|
| 0 | 10 | 32 |

Témoins

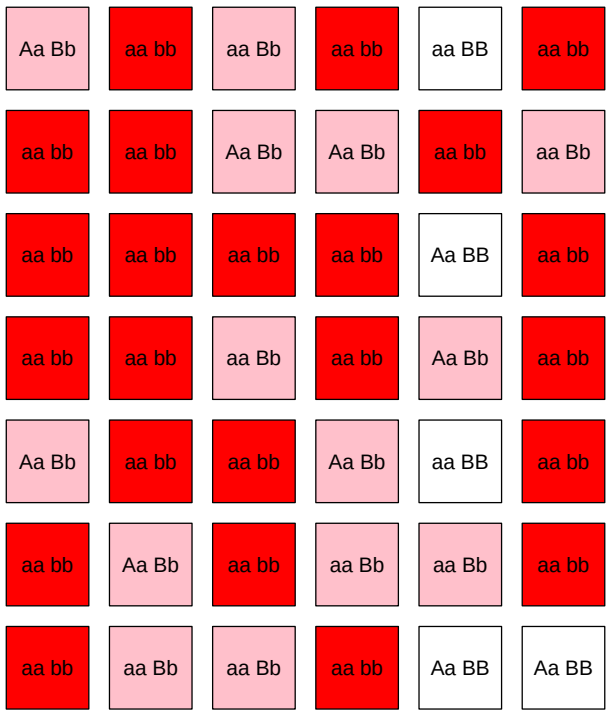



| AA | Aa | aa |
|---|---|---|
| 7 | 22 | 13 |

# Détection d'une association

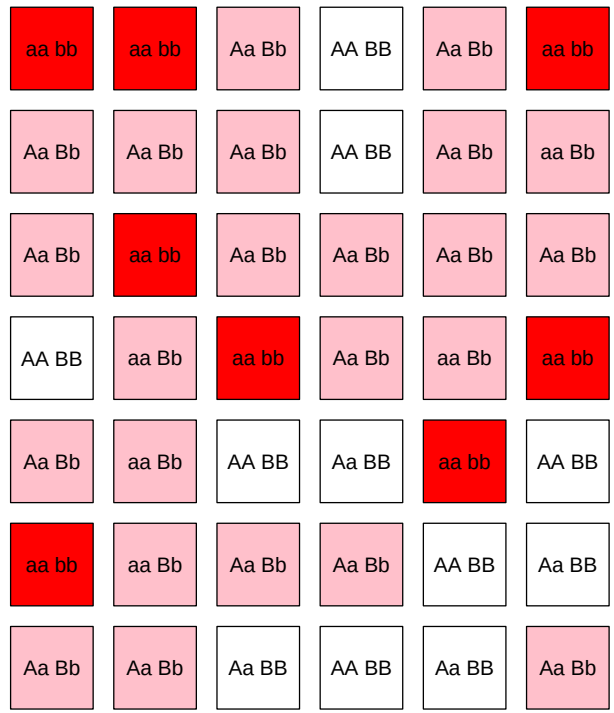
## On observe un SNP corrélé avec le précédent

Cas




| BB | Bb | bb |
|---|---|---|
| 5 | 14 | 23 |

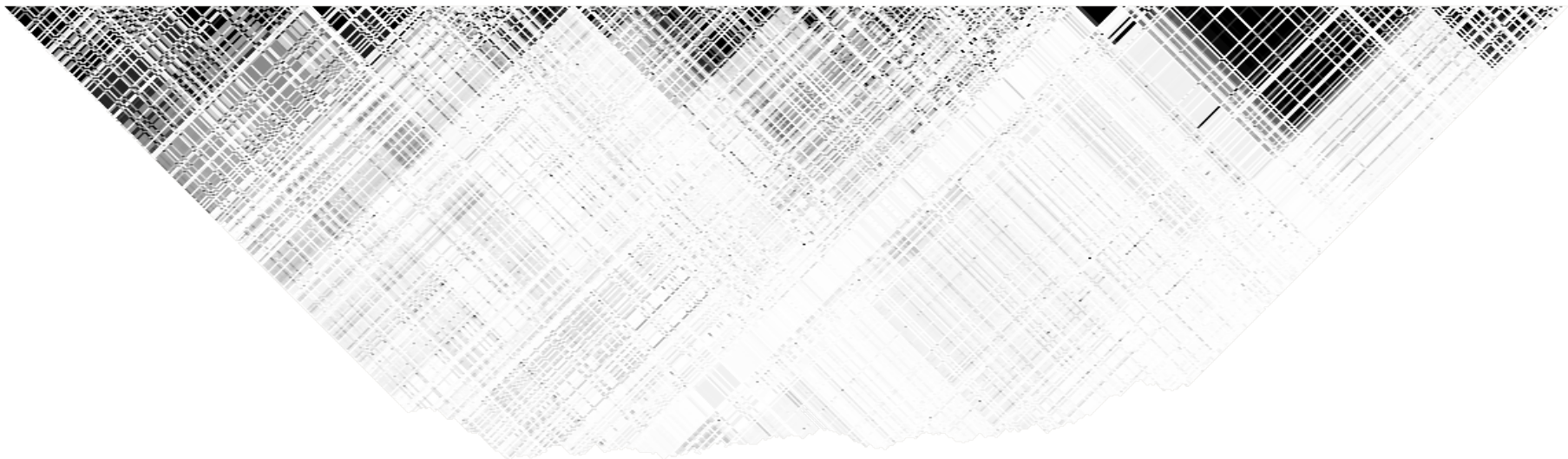
Témoins




| BB | Bb | bb |
|---|---|---|
| 11 | 23 | 8 |

# Motif de déséquilibre de liaison sur le génome humain

Cette figure montre le déséquilibre de liaison entre 970 SNP répartis sur 1 Mb sur le chr 5 (hapmap CEU).



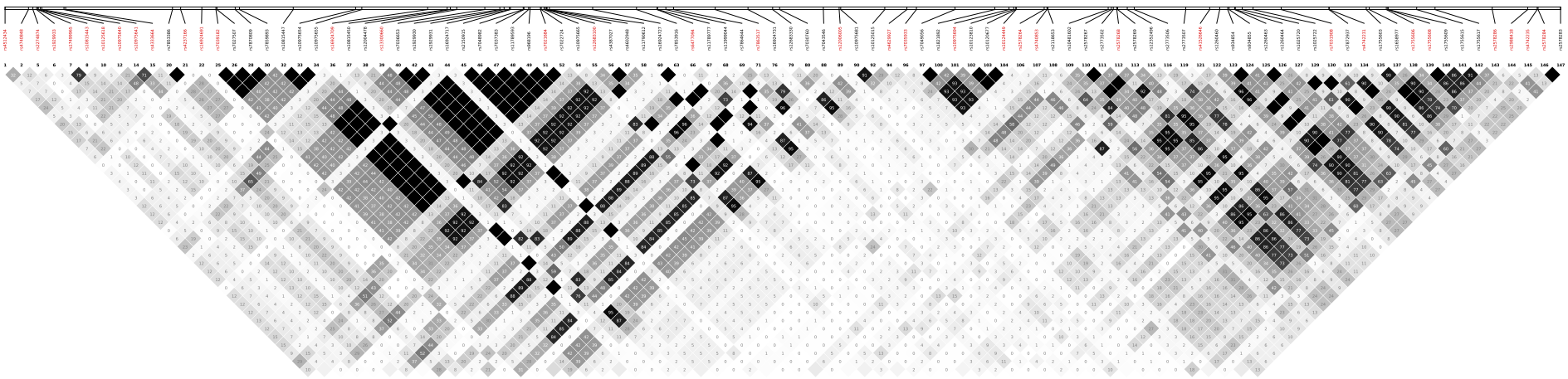
Ce motif en blocs permet l'utilisation de tag-SNP dans les études d'associations.

Ce motif est lié à l'existence de points chauds *(hotspots)* de recombinaison

# Tag-SNP

Pas besoin de tester l'association avec tous les SNP du gène : grâce au DL, un SNP peut être corrélé avec plusieurs autres.

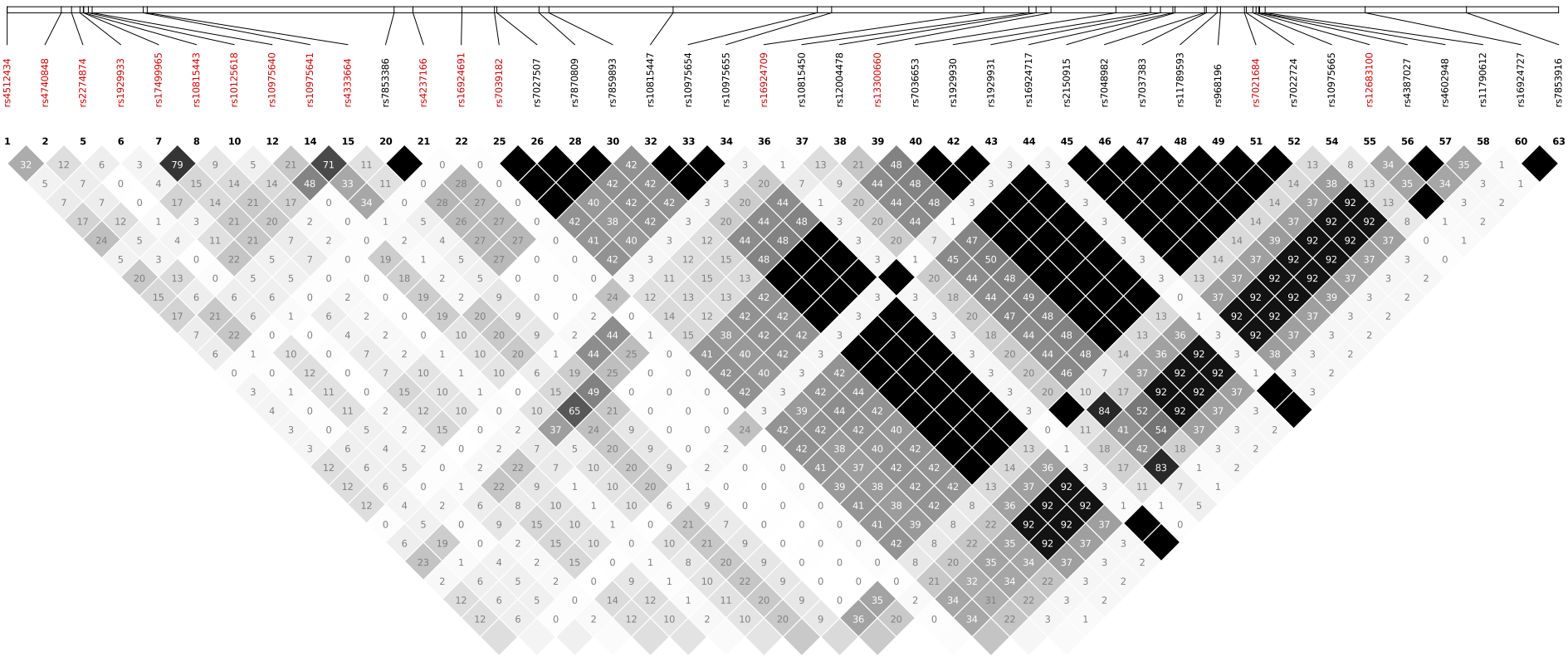
La base de donnée Hapmap donne 96 SNP dans GLDC (glycine dehydrogenase) dans les populations européennes. 38 SNP (en rouge) sont suffisants pour « taguer » tous ces SNP avec $r^2 > 0.8$.



Données Hapmap + logiciel Haploview

# Tag-SNP

Un zoom sur le graphe précédent.

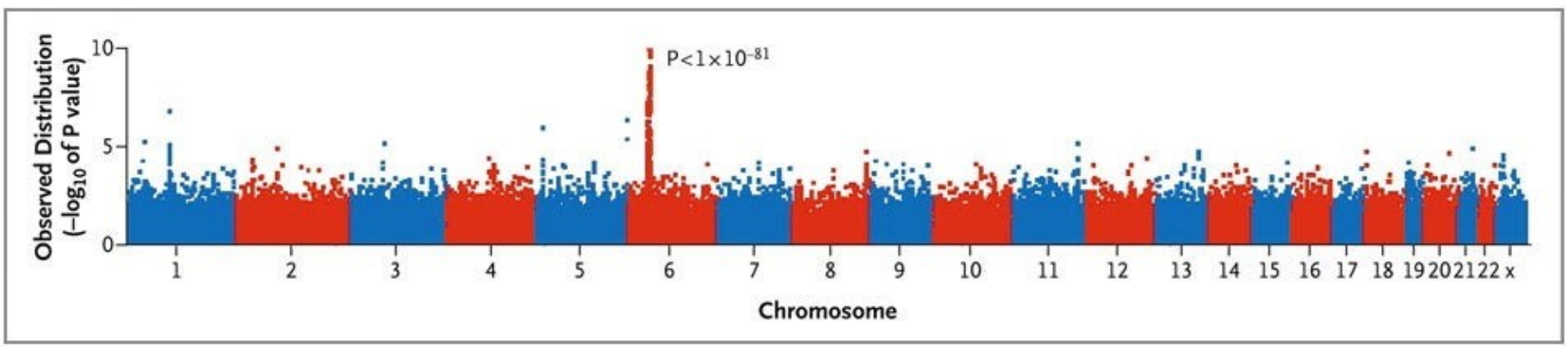


# Études d'association pangénomiques

On prend tout le génome comme candidat

- Grâce aux données de HapMap, on peut choisir des tag-SNP pour le génome entier

  Des SNP arrays commerciaux, créés pour ce but, sont disponibles
- Tests multiples : seuil de significativité $5 \cdot 10^{-8}$
- De grands échantillons sont nécessaires (de 1000 à plusieurs dizaines de milliers de cas et témoins...)
- ☞ Problèmes de stratification de population

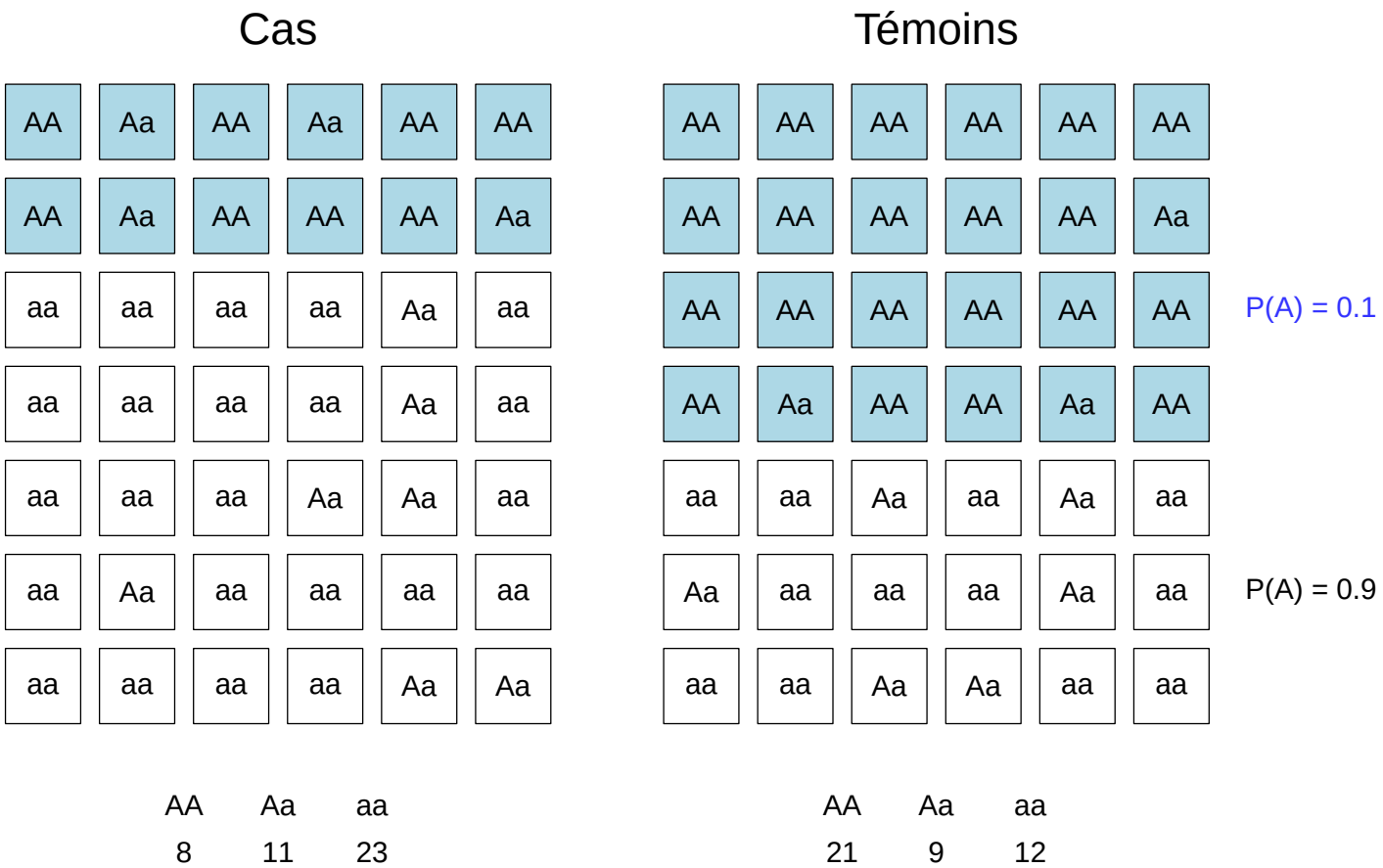



Manhattan plot, IMSGC Lancet 2007

En utilisant des panels de référence constitués d'individus dont la totalité du génome a été séquencée (1000 genomes project), on peut également inférer avec une précision satisfaisante le génotypes en des SNPs qui ne sont pas présents sur le SNP array.

Note : le seuil de significativité de $5 \cdot 10^{-8}$ est considéré comme correct pour toute association de génome entier, avec ou sans imputation

# La stratification de population crée de fausses associations



# Stratification de population

La population européenne n'est pas panmictique.

- Appariemment préférentiel par pays
- Chaque pays divisé en régions
- Possibilité d'homogamie (taille ? HLA ?)
- Sous-population de migrants

☞ Patchwork de sous-populations, dont aucune n'est totalement isolée.

Les différences entre ces sous-populations peuvent persister longtemps, et sont constament renouvelées (par exemple par la dérive génétique).

On parle de « stratification de population ».

# Stratification de population

## Utiliser l'ACP pour visualiser la stratification

On considère des individus genotypés en $p$ SNP di-alléliques. Les génotypes sont codés $0, 1, 2$ (homozygote pour l'allèle de référence, hétérozygote, homozygote pour l'allèle alternatif).

Le génotype d'un individus est récapitulé par un long vecteur du genre

$$(a_1, a_2, \ldots, a_p) = (0, 1, 0, 0, 1, 2, 2, 0, \ldots)$$

On « réduit la dimension » en calculant à partir de ce long vecteur un petit nombre de « scores », calculé comme ceci : formule du style

$$S = u_1 a_1 + u_2 a_2 + \cdots + u_p a_p$$

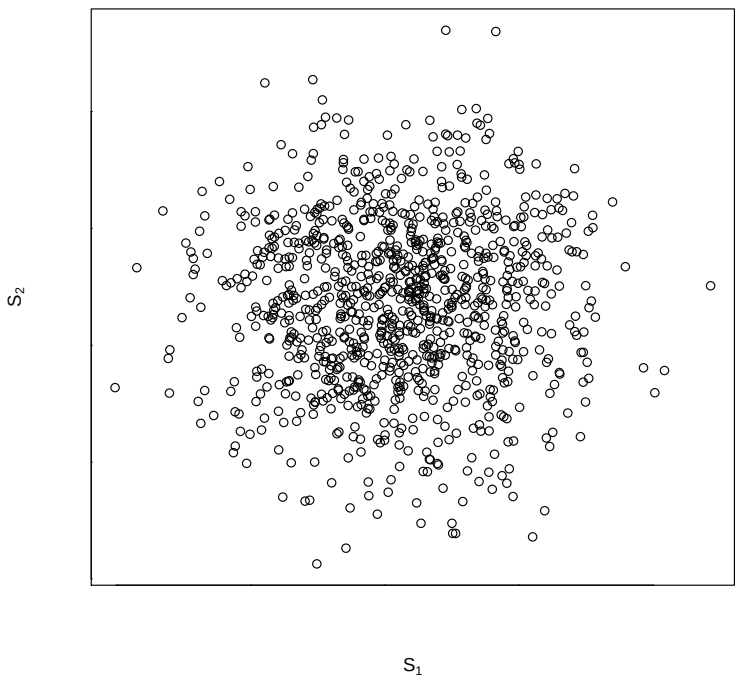
où les valeurs des $u_k$ sont les « charges » des SNP.

# Stratification de population

## Utiliser l'ACP pour visualiser la stratification

Ces scores vont permettre de « résumer » l'ensemble du génome de l'individu.

Calculons deux scores $S_1$ and $S_2$ pour chaque individu et traçons les points de coordonnées $(S_1, S_2)$. Si on choisi les charges au hasard, on obtient quelque chose comme ça.

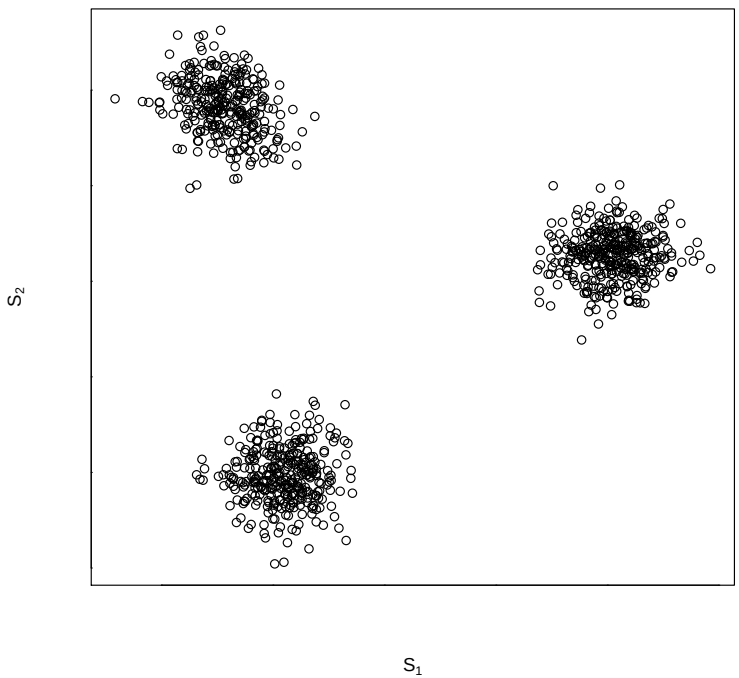


# Stratification de population

## Utiliser l'ACP pour visualiser la stratification

Maintenant nous choisissons les charges qui « maximisent la dispersion » des points.

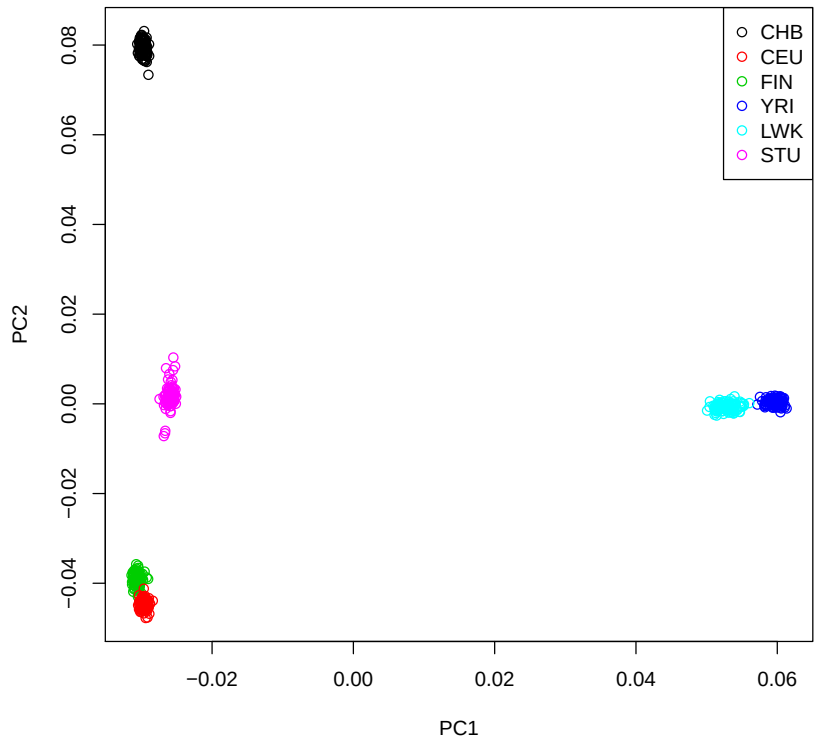
Ici notre population se sépare en trois populations distinctes.



# Stratification de population

## Utiliser l'ACP pour visualiser la stratification

Voici le résultat en utilisant des populations issues de continents différents

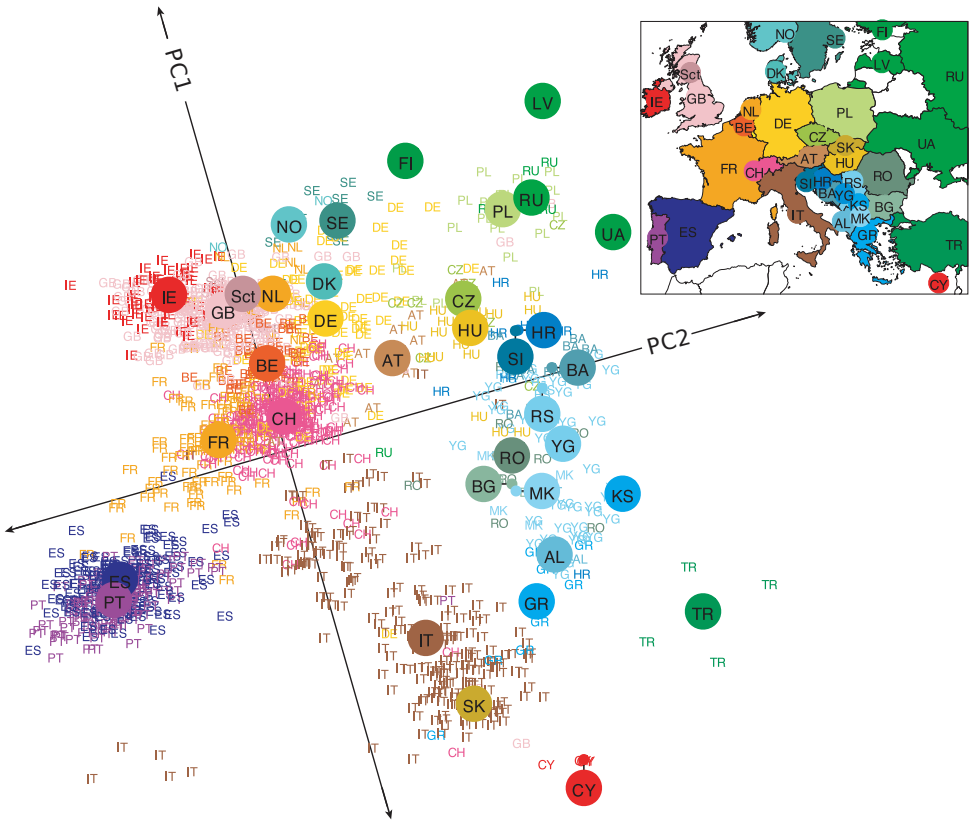



CHB = Han Chinese in Beijing, CEU = Central Europeans in Utah, FIN = Finns in Finland
YRI = Yoruba in Idaban, LWK = Luhya in Webuye, Kenya, STU = Sri Lankan Tamil in UK

# Stratification de population

## Utiliser l'ACP pour visualiser la stratification

Et voici le résultat en utilisant un échantillon d'européens.




Novembre 2008, Genes mirror geography within Europe

# Prendre en compte la stratification de population

- On peut tester l'équilibre de Hardy Weinberg pour supprimer les SNP les plus affectés par la stratification (supprime également les SNP mal génotypés)
  - On parle d'« équilibre de Hardy-Weinberg » quand les trois génotypes possibles en un SNP sont dans des proportions

| AA | Aa | aa |
|---|---|---|
| $p^2$ | $2pq$ | $q^2$ |

    où $p$ est la fréquence de l'allèle A, $q$ la fréquence de l'allèle a ($p + q = 1$).
  - Ça doit être le cas (aux fluctuations aléatoire près) si la population est *panmictique*, si le locus génétique n'est pas sous sélection, etc.

# Prendre en compte la stratification de population

- On utilise les composantes principales (CP) comme covariables dans le test

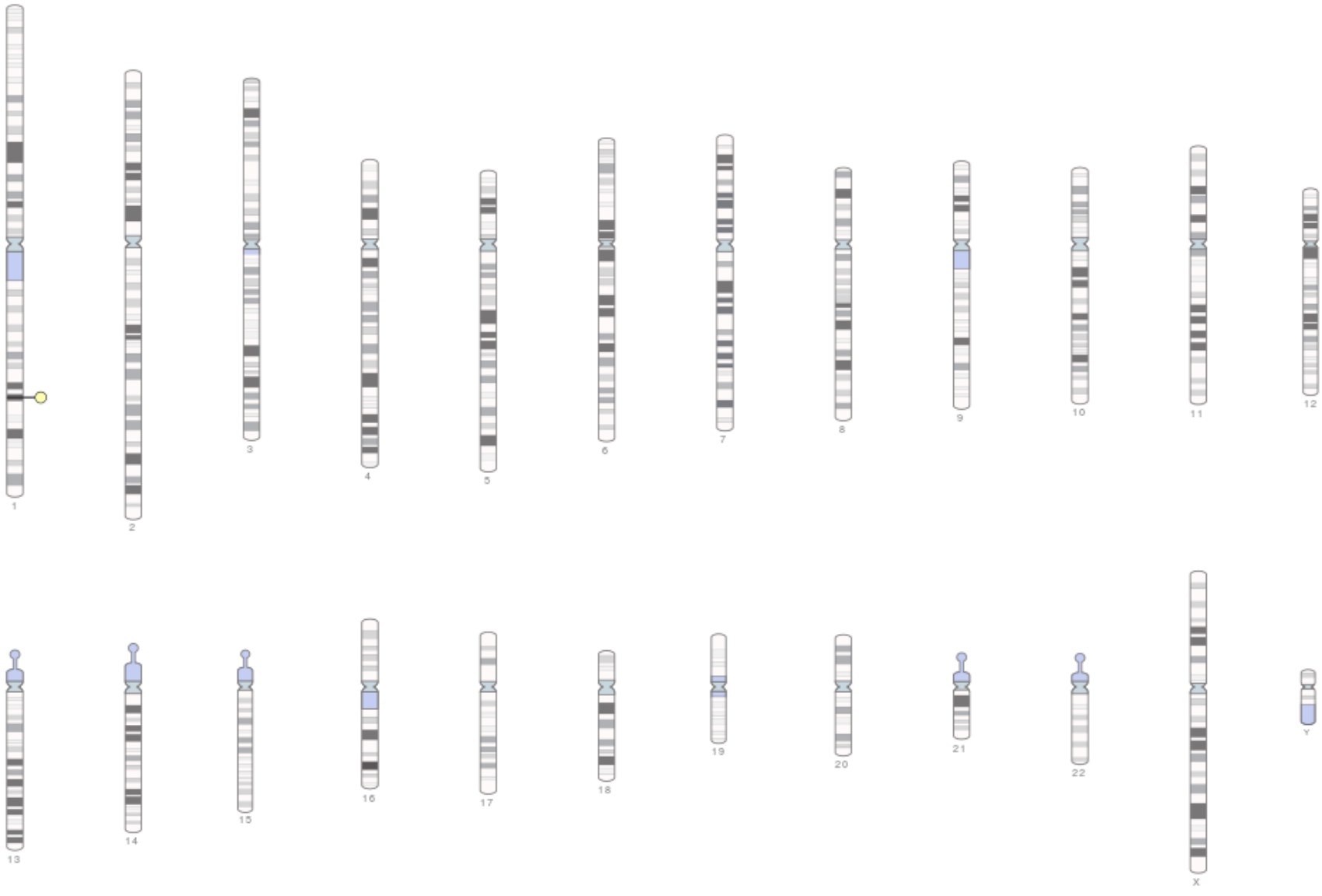
$$\text{logit } P(Aff) = \alpha + \beta SNP + \gamma_1 CP_1 + \gamma_2 CP_2 + \cdots$$

  - Nous avons vu que les CP reflètent la population d'origine
  - Même démarche que l'ajustement sur le sexe, l'âge, l'environnement...

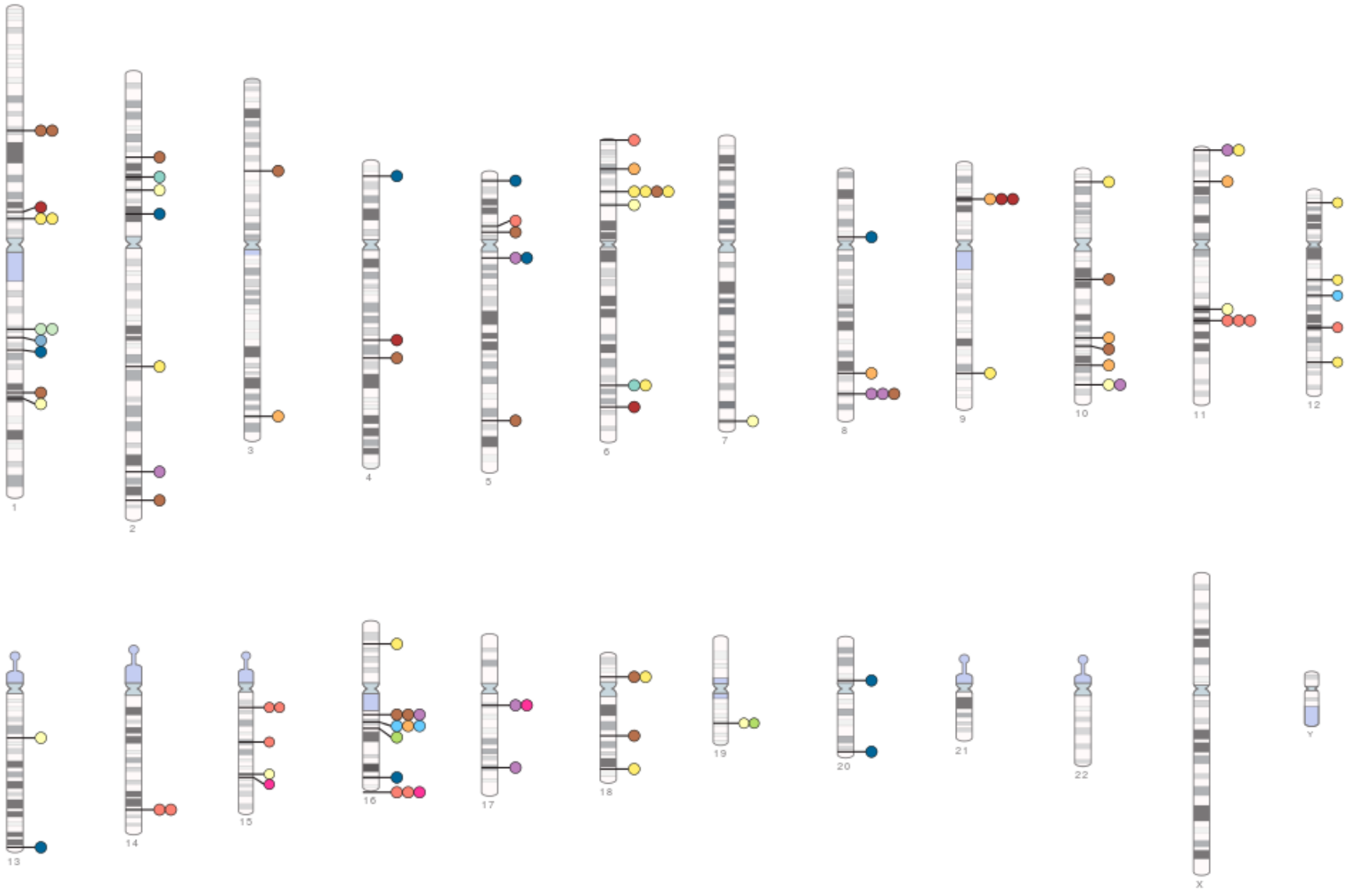
- Utilisation de données familiales
  - Transmission Disequilibrium Test (TDT), avec des trios
  - Family Based Association Test (FBAT), avec des familles nucléaires

  Méthodologie malheureusement ( ?) passée de mode

# Résultats des études GWAS (fin 2005)



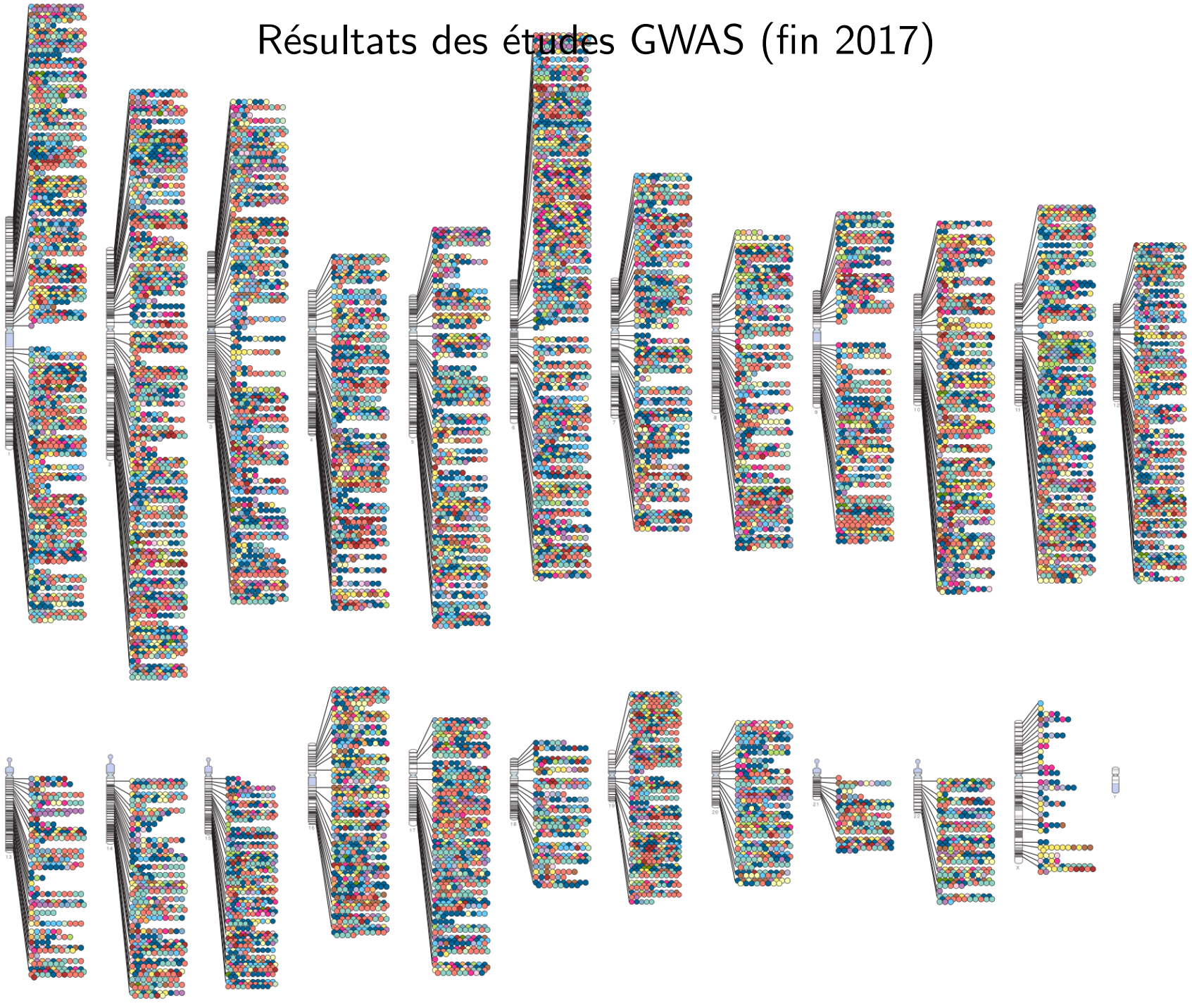
# Résultats des études GWAS (fin 2007)



# Résultats des études GWAS (fin 2009)

# Résultats des études GWAS (fin 2011)

# Résultats des études GWAS (fin 2017)



# Petit bilan des études GWAS...

- Environ 10 000 SNPs associés dans environ 2000 études répertoriées (NHGRI GWAS Catalog)
- La plupart du temps, les risques relatifs sont faibles
- ☞ L'utilité pour la médecine prédictive (ou « médecine personnalisée ») reste débattue
- Meilleure compréhension de l'étiologie dans certains cas

# Analyse des variants rares

- « Héritabilité manquante » : les variants découverts dans les études d'association pan-génomiques (GWAS) ne semblent pas suffire à expliquer toute l'héritabilité de certains traits quantitatifs ou l'agrégation familiale observée dans les maladies complexes
  - ☞ mais celles-ci sont conçues pour détecter les variants fréquents
- Arrivée sur le marché du séquençage haut-débit (exome seq)
  - ☞ disponibilité des données
- L'analyse SNP par SNP pratiquée dans les GWAS n'a pas la puissance statistique nécessaire à détecter les variants rares
  - ☞ développement de tests appropriés

# Analyse des variants rares

- Méthodes permettant l'analyse « gène par gène » (ou région par région) : un gène = un test.

  ☞ moins de tests : regain de puissance (environ 20 000 gènes sur le génome humain)

  Deux familles de méthodes se sont développées :

- tests de fardeau, ou *burden tests*

  On calcule un « fardeau » (p. ex. nb de variants rares dans le gène) et on compare la distribution des fardeaux entre cas et témoins

- tests « de variance » : SKAT & Co

  Postulent un modèle l'effet pour l'effet de tous les variants du gène

Le séquençage de tout le génome est maintenant à l'ordre du jour.

# Choses dont je n'ai pas parlées

## et dont je peux parler brièvement maintenant

Quelques points dans les angles morts

- les scores polygéniques
- les *summary statistics*
- les biothèques *(biobanks)*
- la randomisation mendélienne
- ...

...c'est fini !