La cartographie génétique

Quant à elle, mesure la probabilité dans une famille qu'ont deux gènes de ségréger ensemble au cours de la méiose, décrivant en cela le comportement des gènes en méiose plutôt qu'une localisation physique. Dans la cartographie génétique, la fréquence de recombinaisons méiotiques est utilisée pour estimer les distances entre les marqueurs. La distance génétique est une mesure statistique estimée en centi-Morgan (cM).

L'analyse de liaison génétique est une méthode permettant de localiser les gènes par l'étude de leur co-ségrégation (c'est-à-dire de leur liaison) sur un chromosome au cours de la méiose. La liaison génétique est extrêmement importante en génétique médicale, car pour la grande majorité des gènes responsables des maladies génétiques, ni leur base biochimique, ni leur base moléculaire ne sont identifiées.

1-Les recombinants et les non recombinants

Les paires de chromosomes homologues s'apparient au cours de la méiose I et subissent un certain nombre de recombinaisons; échangeant des sections homologues par crossing-over (CO) et créant de nouvelles combinaisons d'allèles dans les produits de la méiose. Le crossing-over a lieu au moment de la méiose où il existe quatre chromatides (tétrade). Cependant, chaque crossing-over implique seulement 2 des 4 chromatides. Pour chaque recombinaison, il existe quatre produits: deux qui sont recombinants (R) et deux non recombinants (NR) (figure 15).

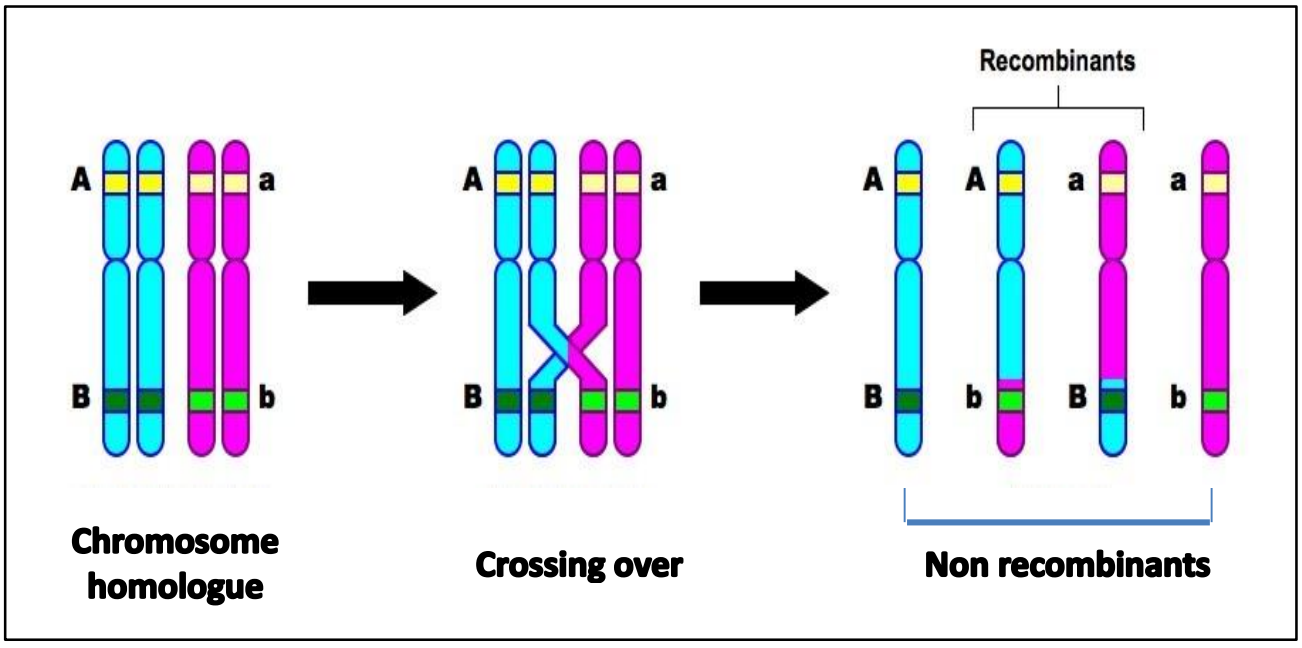

Figure 15 : les recombinants et les non recombinants

2-Ségrégation des gènes

Il y a trois façons possibles pour une paire d'allèle de ségréger en méiose:

2-1- Premier cas de ségrégation (gènes indépendants)

Les allèles situés sur différents chromosomes se combinent de façon indépendante en méiose. Pour deux loci A et B avec les allèles A/a, et B/b, nous observons quatre types de gamètes en proportions égales (25%), les deux gènes seront considérés comme non liés et la ségrégation devient de type indépendant. On dit que deux gènes sont indépendants quand leur distance génétique est supérieure de 50 cM (D ≥ 50 cM, P=R=50%). Ceci peut signifier que les deux gènes sont sur des chromosomes différents ou très éloignés sur le même chromosome (figure 16).

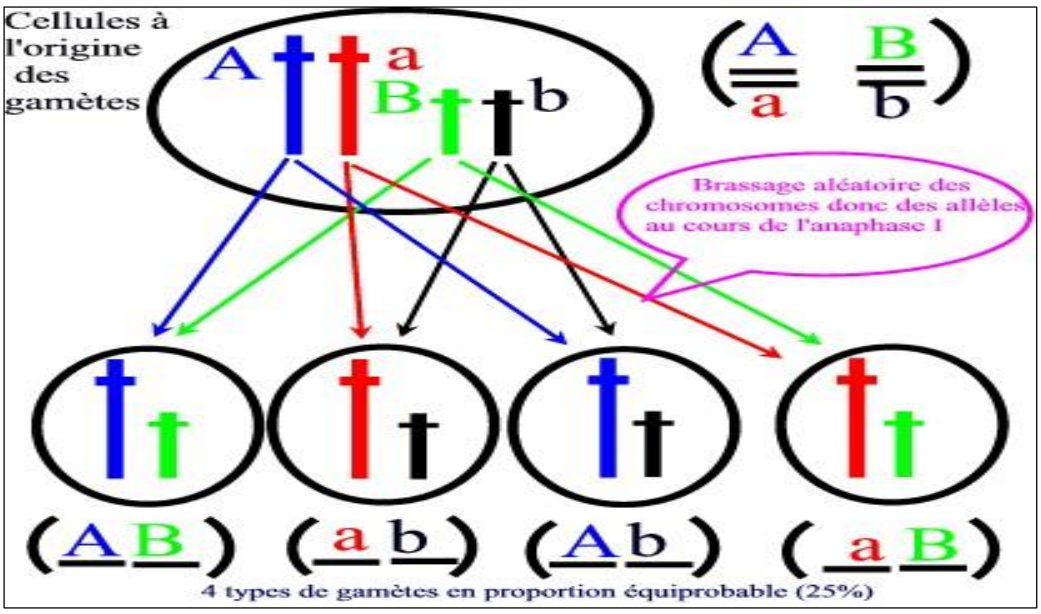

Figure 16 : les gènes indépendants

2-2- Deuxième cas de ségrégation (liaison génétique)

Les gènes situés côte à côte sur le même chromosome sont transmis pratiquement toujours ensemble. En conséquence, seuls deux des combinaisons alléliques possibles vont être observées dans la descendance; lesquelles dépendront de la combinaison particulière des chromosomes parentaux AB et ab. L'assortissement ne sera pas indépendant et aucune recombinaison ne sera observée. La probabilité d'avoir un crossing-over entre les deux gènes tend à s'annuler lorsqu'ils sont proches (D < 10 cM), il s'agit d'une liaison totale (pas d'échange). (figure 17).

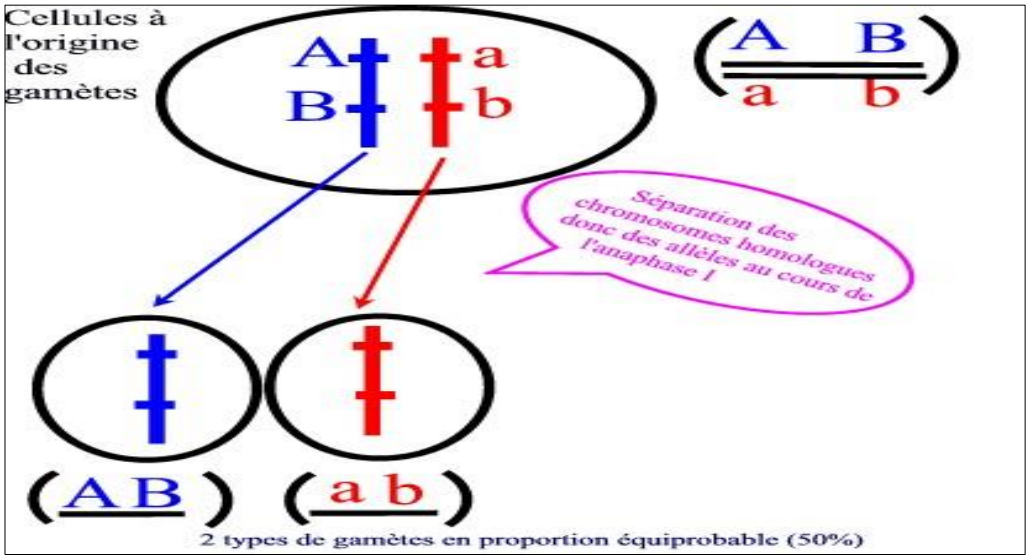

Figure 17 : liaison absolue des gènes

2-3- Troisième cas de ségrégation (gènes liés)

Entre ces deux extrêmes, les allèles de deux loci localisés à une certaine distance l'un de l'autre sur le même chromosome tendent à être transmises ensemble, au moins qu'une recombinaison au cours de la méiose ne crée une nouvelle combinaison. Dans ce cas, nous observons quatre combinaisons d'allèles dans la descendance. Les proportions relatives de ces combinaisons dépendront de fréquence de recombinaison entre les deux loci. On dit que deux gènes sont liés quand leur distance est inférieure à 50cM (P > 50% et R < 50 %). Ceci signifie que les gènes sont proches sur le même chromosome (figure 18).

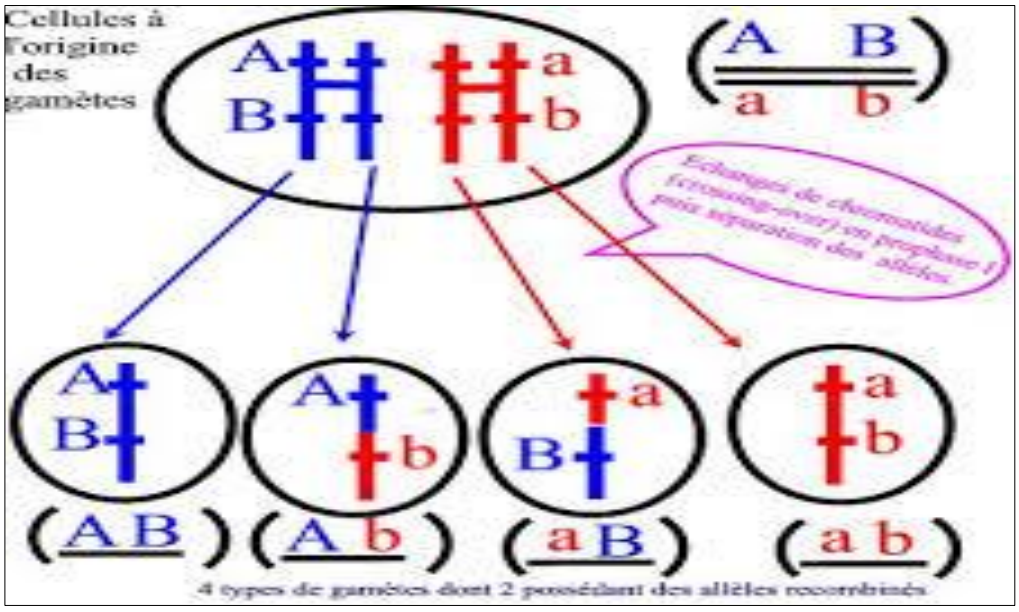

Figure 18 : les gènes liés

3-La liaison génétique

La liaison génétique peut être définie comme la tendance pour deux allèles situés proches l'un de l'autre sur le même chromosome **d'être transmis ensemble comme une seule unité** au cours de la méiose. La force de liaison peut alors être utilisée comme unité de mesure permettant d'estimer la proximité de deux loci distincts. Cette unité de mesure est un reflet de la distance physique.

4-La distance génétique

La distance génétique (D) est la distance qui sépare deux loci sur un chromosome. La distance entre deux loci peut être déduite en estimant la **fréquence de recombinaisons** survenant dans les familles. 1cM est approximativement égal à une fréquence de recombinaison de 1%. La distance génétique est proportionnelle à la fréquence de recombinaison entre deux gènes liés. L'unité de la distance génétique est le centi Morgan (cM).

D (A-B) = fréquence des gamètes recombinants pour les gènes A et B X 100

$$D = \frac{\text{Nombre de recombinants}}{\text{Nombre total de descendants}} \times 100$$

5- L'équivalence entre la distance génétique et la distance physique

1 cM est une unité de longueur génétique sur laquelle on observe une recombinaison une fois sur 100. La longueur totale du génome a été estimée à 3000 cM sur la base du nombre de chiasmas observés au cours de la méiose I dans la spermatogénèse. Puisque le génome haploïde (gamète) correspond à une longueur physique d'environ 3×10^9 pb donc 1cM est équivalent à environ 1 million de pb

$1cM = 10^6$ pb

Selon cette estimation, il est suggéré qu'un chromosome moyen est long de 100 à 300 cM. On s'attend par conséquent à observer en moyenne 1 à 3 recombinaisons par

chromosome, par méiose. Cependant, il ne s'agit que d'une relation très approximative. Plusieurs facteurs sont susceptibles d'influencer les taux de CO:

- les CO sont environ 1,5 fois plus fréquents au cours de l'ovogenèse que pendant la spermatogenèse
- les CO tendent à être plus fréquents à proximité des télomères des chromosomes que dans les régions des centromères
- certaines régions chromosomiques montrent des taux de CO particulièrement plus élevés, ces régions sont qualifiées de points chauds de recombinaison (hot spot); les séquences Alu semblent particulièrement sujettes à des CO.

6-Détection et mesure de la liaison génétique

Il y a **deux conditions importantes** pour procéder à l'analyse de liaison:

- une **famille informative** pour les loci qui sont considérés, la nécessité qu'un parent soit hétérozygote à la fois pour le locus morbide et le locus marqueur
- **la connaissance de la phase :** la disposition particulière des allèles chez les parents aux deux loci doit être connue ou doit pouvoir être déterminée.

NB : La phase de liaison est la façon selon laquelle se dispose un couple de gènes liés sur une paire chromosomique homologue. Liaison en couplage (cis) et liaison en répulsion (trans)

Il est évident que pour la détection d'une liaison génétique, de très grandes familles sont plus utiles que de petites familles; de même les familles de trois générations sont plus utiles que les familles de 2 générations, mais dans les études de liaison de maladies génétiques, de telles familles ne sont pas toujours disponibles. Il existe **deux approches** parallèles et apparentées pour **l'étude de la liaison en génétique** humaine:

- développer une carte de liaison génétique détaillée avec une batterie de très grandes **familles de malades**, de trois générations. L'objectif de cet effort est de mesurer de façon aussi exacte et aussi précise les relations de liaison entre plusieurs loci informatifs, en particulier des marqueurs
- établir la liaison de loci chez des **familles normales** et ne manifestant aucune maladie génétique connu. Il est important d'avoir à sa disposition une carte génétique normale détaillée pour comparaison.

7-Les marqueurs génétiques

Une carte génétique est facilement construite chez la levure, la drosophile et la souris, mais chez l'homme, les familles humaines dans lesquelles deux maladies ségrègent, sont extrêmement rares et même lorsque de telles familles peuvent être trouvées, elles peuvent avoir très peu d'enfants ou être inexploitables. Pour cette raison, la plupart des études de cartographie génétique humaines utilisent **des marqueurs génétiques.** Un marqueur génétique est un gène ou une séquence polymorphe d'ADN (variations identifiables) aisément détectable grâce à un emplacement connu sur un chromosome.

7-1-Les types de marqueurs

7-1-1-Les marqueurs morphologiques

Ils sont basés uniquement sur des gènes cartographiés à partir de l'analyse des caractères morphologiques ou phénotypiques (couleur des poils, capacité à pousser en milieu carencé, etc…). Ils sont faciles à mesurer mais peu polymorphes, rarement co-dominants et ambigus.

7-1-2-Les marqueurs moléculaires

Il s'agit de polymorphismes au niveau de l'ADN (SNP, microsatellites, RFLP,…). Ils sont nombreux, très polymorphes, co-dominants, non ambigus et requièrent la biologie moléculaire pour être détectés.

7-2-Les propriétés des marqueurs

Qu'est-ce qu'un bon marqueur dans une optique de cartographie? Les marqueurs doivent présenter **plusieurs propriétés**, pour pouvoir être utilisés dans une analyse de cartographie génétique. Les marqueurs doivent être :

- **polymorphes** présenter différents allèles identifiables. Le locus comprend de nombreux allèles pour s'assurer que la plupart des parents seront hétérozygotes pour le locus marqueur
- **co-dominants** les homozygotes doivent pouvoir être distingués des hétérozygotes. Les 2 allèles sont détectables chez les hétérozygotes. Cette caractéristique rend plus facile la détermination de la phase de liaison.
- **nombreux** afin qu'une liaison étroite avec le gène pathologique soit plus probable. Chaque chromosome est aujourd'hui saturé de marqueur. Plusieurs milliers de marqueurs ont désormais été identifiés dans tout le génome.

- **à site unique** dans le génome

7-3-Exemples de marqueurs moléculaires

Il existe différentes sortes de marqueurs :

- **RFLP**, Restriction fragment length polymorphisme, polymorphisme de longueur des fragments de restriction
- **AFLP**, Amplified fragment length polymorphism, polymorphisme de longueur des fragments amplifiés
- **RAPD**, Random amplified polymorphic DNA, amplification aléatoire d'ADN polymorphe
- **SNP**, Single nucleotide polymorphism, polymorphisme simple base
- **SSCP**, Single strand conformation polymorphism, polymorphisme de conformation des simples brins
- **EST**, Expressed sequence tags, marqueur de séquence exprimée
- **VNTR**, Variable number of tandem repeats, nombre variable de répétition en tendem

7-4-Les marqueurs génétiques et la cartographie génétique

Il existe deux approches générales dans l'utilisation des marqueurs dans la cartographie génétiques:

- cartographie par **marqueur-maladie,** utilisée pour localiser les gènes de maladies ;
- cartographie par **marqueur-marqueur**, celles-ci aide la cartographie marqueur-maladie et aide à établir les correspondances entre carte génétique et physique en localisant les mêmes locus sur chacun des types de cartes.

Les marqueurs les plus utilisés en cartographie sont les SNP, RFLP et les VNTR

7-4-1- SNP

C'est la variation (polymorphisme) d'une seule paire de bases du génome, entre individus d'une même espèce. Ces variations sont très fréquentes, environ une paire de base sur deux milles dans le génome humain.

Les SNP représentent 90 % de l'ensemble des variations génétiques humaines, ils sont présents toutes les cent à trois cents paires de bases en moyenne dans le génome humain. Deux SNP sur trois substituent la cytosine avec la thymine.

Les SNP peuvent se retrouver au sein de régions codantes de gènes (exons), de régions non codantes de gènes (introns), ou de régions intergéniques (entre les gènes).

7-4-2-RFLP

L'ADN d'un individu ou d'une cellule est d'abord extrait et purifié. L'ADN est ensuite coupé en fragments de restriction par une enzyme de restriction, les fragments d'ADN ainsi obtenus, nommés fragments de restriction, sont ensuite séparés selon leur longueur par électrophorèse sur gel d'agarose. Le gel obtenu est ensuite analysé par Southern blot et révélé avec une ou plusieurs sondes (figure 19).

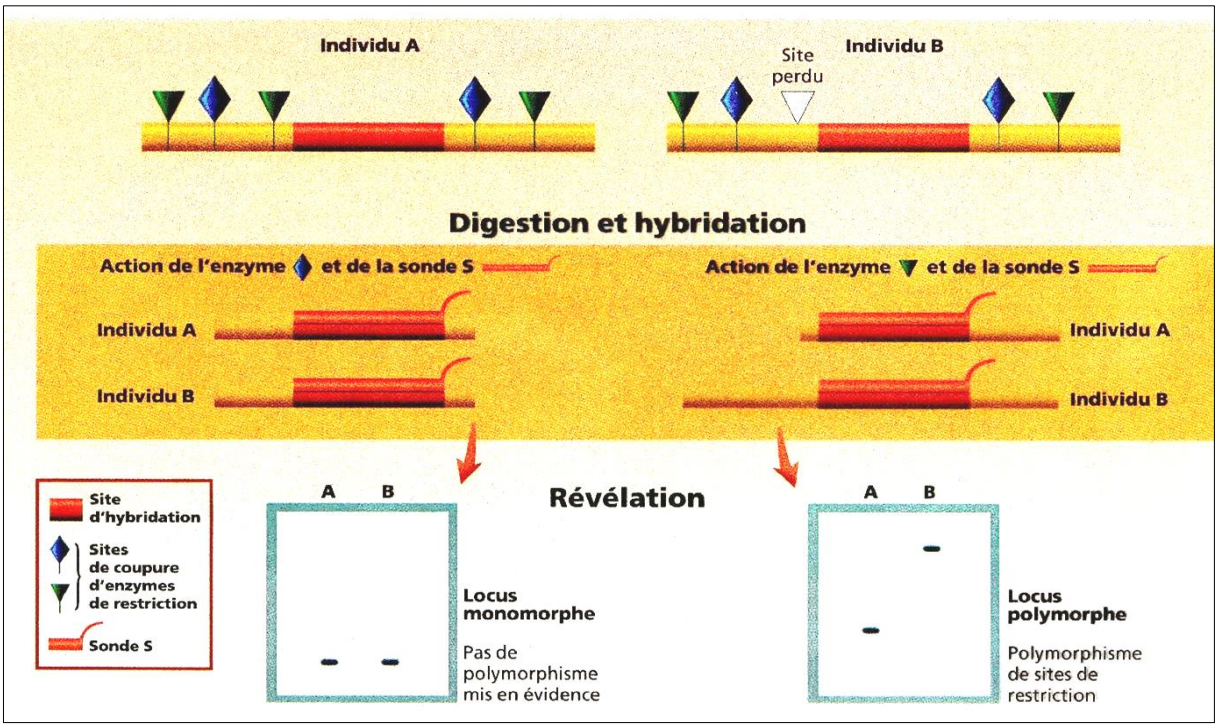

Figure 19 : les marqueurs RFLP

7-4-3-VNTR

Séquences localisées répétées en tandem, il en existe différentes catégories :

- les **microsatellites** (SSR) : sont des séquences de 1 à 5 paires de bases répétées un grand nombre de fois, des répétitions longues de plusieurs Kb, il y a plus de 10 000 zones de répétition dans le génome, le plus fréquent et le plus utilisé sont les dinucléotides (CA)/(GT). La méthode de génotypage est l'amplification et le séquençage du produit de PCR. Ils présentent une hétérozygotie souvent supérieure à 70% (figure 20).

- Les **minisatellites** sont des séquences de 15 à 25 pb répétées un grand nombre de fois (1000 à 2000 fois), répétions de séquences inférieures à 150 pb. Ils sont télomériques et très polymorphes. Explorées par technique de southern, leur hétérozygotie est supérieure à 90 %.
- **Les grands blocs d'ADN satellite** (environ 10 % du génome humain) : c'est des blocs allant jusqu'à une dizaine de mégabases, localisés très majoritairement au niveau des centromères et des télomères.

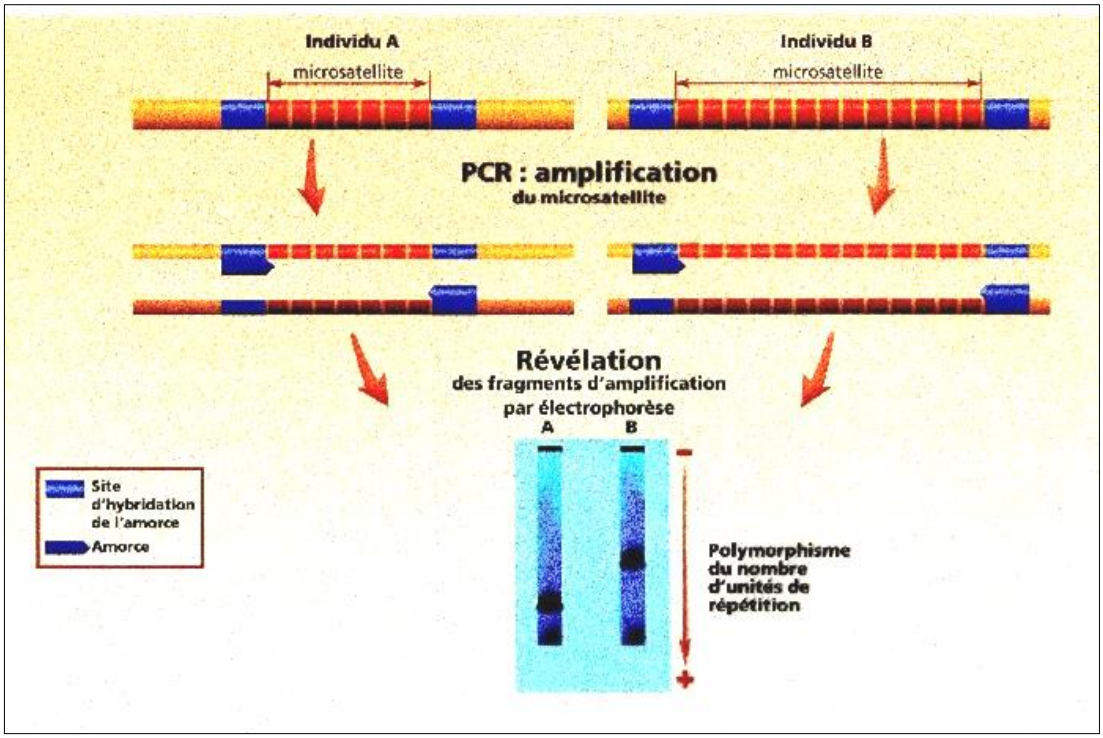

Figure 20 : les marqueurs microsatellites

8-Le déséquilibre de liaison

Le déséquilibre de liaison comme **une association non aléatoire** de deux ou plusieurs loci, cela signifie que les gènes appartenant à des loci différents ne sont pas associés au hasard dans la population, certaines combinaisons sont moins fréquentes, d'autres sont plus fréquentes que ne le voudrait le hasard si on avait une association aléatoire de ces différents gènes, c'est-à-dire **il y a association préférentielle entre les deux allèles.**

On mesure la force du déséquilibre entre deux loci à l'aide du coefficient de déséquilibre de liaison **(D)**. Soit deux locus possédant respectivement les allèles A/a et B/b, le coefficient de déséquilibre de liaison correspond à la différence entre la proportion

d'haplotype A/B ou a/b **observée** et la fréquence **théorique** calculée qui est le produit des fréquences.

- En **équilibre de liaison**, la fréquence théorique calculée est le produit des fréquences géniques: $\mathbf{P_{AB} = P_A \times P_B}$.
- Dans le cas contraire c'est-à-dire, s'il n y a pas d'indépendance statistique donc il existe un **déséquilibre de liaison**, on ne peut plus écrire cette égalité, puisque la fréquence du gamète AB sera soit supérieure, soit inférieure aux produits des fréquences des deux gamètes A et B : $\mathbf{P_{AB} = P_A \times P_B \pm (D)}$

$$\mathbf{D = P_{AB} - (P_A \times P_B)}$$

Les recombinaisons génétiques réduisent à chaque génération la valeur du déséquilibre de liaison, mais si les gènes sont étroitement liés, l'estimation vers l'équilibre est très lente. De nombreux mécanismes peuvent être à l'origine du déséquilibre de liaison:

- le mélange des populations ;
- les mutations ;
- et la sélection naturelle.

8-1- Le déséquilibre de liaison et la cartographie génétique

Le concept du déséquilibre de liaison est très utile pour la cartographie génétique, ceci s'explique par le fait que si on trouve une association entre un marqueur et une maladie cela suggère qu'il existe un déséquilibre de liaison entre eux. En d'autres mots, le locus marqueur et le locus de la maladie sont **étroitement liés.**

9-Le LOD score (Z ou logarithm of the odds)

Lors des analyses de liaison génétique, on se demande souvent si deux gènes sont liés, parfois, la réponse est évidente, car la fréquence de recombinants est nettement inférieure à 50%, parfois elle ne l'est pas. Dans ces deux situations, il est cependant utile d'appliquer un test statistique objectif pour confirmer ou infirmer notre intuition.

Les progrès de la cartographie des loci des gènes humains furent très lents au départ pour plusieurs raisons : tout d'abord, il n'est pas possible d'effectuer des unions contrôlées chez les humains et les généticiens durent calculer les fréquences de recombinants à partir des di-hybrides occasionnels apparus par hasard à la suite d'unions humaines. Les croisements équivalents au croisement-test sont extrêmement rares. De plus les unions humaines en général ont de petits nombres de descendants, ce qui rend difficile l'obtention

de données pour calculer des distances génétiques fiables. Une estimation plus fiable peut être réalisée en combinant les résultats des unions identiques. La procédure classique consiste à calculer des Lods Scores.

Le lod Score (Lod signifie les chances) est une valeur déterminée dans le cadre d'une analyse de liaison génétique. Il s'agit d'une mesure statistique, c'est le logarithme du rapport des probabilités entre l'hypothèse proposée (liaison génétique) et l'hypothèse contraire (pas de liaison).

$$\mathbf{Z = Log_{10}\ (\Theta)}\ \frac{\textbf{la vraisemblance de liaison à une fréquence de recombinaison donnée}}{\textbf{la vraisemblance de l'absence de liaison } (\Theta_{max} = 0.5)}$$

- Dans le cas de liaison génétique ⟶ pas de recombinaison ⟶ $\theta = 0$ ⟶ $Z = -\infty$
- Dans le cas de liaison ⟶ il y des recombinaisons ⟶ $0 < \theta < 0.5$
- Dans le cas d'absence de liaison ⟶ $\theta_{max} = 0.5$

Les individus recombinants font chuter le lod score, car beaucoup de gamètes recombinants indiquent que le gène et le marqueur ont peu de chance d'être génétiquement liés.

9-1- Le calcul de Z

Exemple du calcul de Z pour une fraction de recombinaison $\theta = 5\%$.

$\theta = 5\%$ signifient qu'il y a 5 % de recombinaison entre le gène et le marqueur,

$\theta = 5\%$ cela signifie aussi que les deux gènes resteront liés dans 95% des méioses $(1- \theta)$

Calcul du numérateur de la valeur Z (vraisemblance de liaison):

La probabilité de recombinaison est égale $½ \times 0.05 = 0.025$

La probabilité de non-recombinaison est égale à $½ \times 0.95 = 0.47$

Calcul du dénominateur de la valeur Z (vraisemblance de l'absence de liaison)

La probabilité de la non-liaison pour θ_{max} est de $¼ = 0.25$

Donc $Z = \log_{10} [(0.025 \times 0.47) / 0.25]$

9-2- Interprétation des valeurs du Z

Z ≥ 3 : c'est la preuve d'une liaison (la vraisemblance en faveur de la liaison est 1000 fois supérieure à la vraisemblance d'une absence de liaison)

Z ≤ -2 : c'est la preuve d'une absence de liaison (exclue la liaison génétique, les chances de 100 contre 1 contre la liaison). Il faut typer d'autres marqueurs

- **-2 < Z < +3** : le résultat est ambiguë. Il faut ajouter les lod score d'autres familles **$Z = Z_1 + Z_2 + Z_3 + \ldots\ldots\ldots\ldots Z_n$** (figure 21).

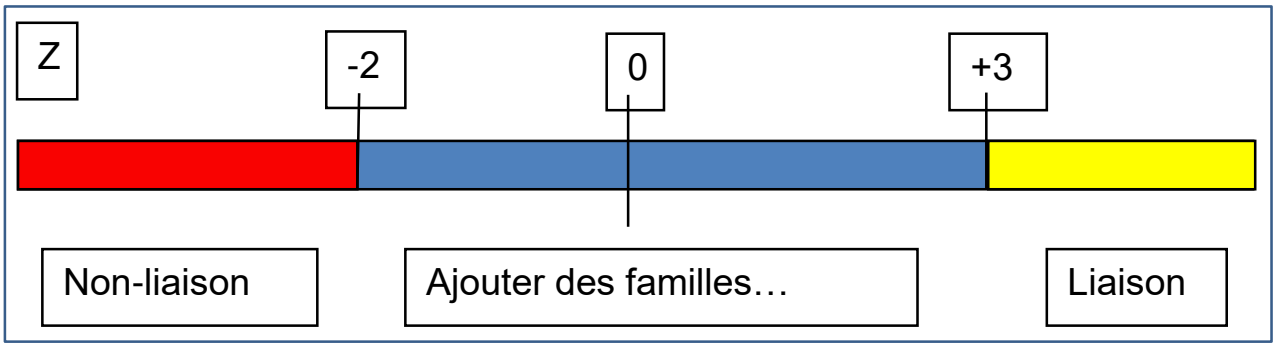

Figure 21 : interprétation des valeurs de Z

NB : Le lod score le plus élevé constitue l'estimation du maximum de vraisemblance de θ, c'est-à-dire qu'il représente la distance la plus probable entre deux loci analysés

Z >>> ⟶ θ = distance entre les deux loci