Etablissement des cartes génétiques

D. VAIMAN

INRA, Laboratoire de Génétique Biochimique et de Cytogénétique, 78352 Jouy-en-Josas cedex

e-mail : daniel.vaiman@biotec.jouy.inra.fr

Résumé. *La construction de cartes génétiques passe par l'analyse de la ségrégation de marqueurs génétiques (polymorphismes de l'ADN, des protéines ou à effets visibles) dans des familles de référence. La distance génétique (exprimée en centimorgan) est proportionnelle au taux de recombinaison existant entre marqueurs polymorphes portés par le même chromosome, au moins pour des distances faibles, observation qui constitue la base de l'établissement des cartes. Quand les distances deviennent très petites (inférieures à 1 cM), l'analyse génétique peut faire appel à l'étude d'haplotypes de marqueurs dans les régions concernées ou à l'analyse du déséquilibre de liaison dans la population.*

La découverte des ordres de grandeur relatifs des génomes des organismes remonte à plus de trente ans. La raison majeure pour laquelle cette donnée fait partie de notre patrimoine scientifique depuis si longtemps est sa relative simplicité d'obtention. En effet, il suffit de compter des cellules, puis d'extraire leur ADN. La quantité d'ADN est alors évaluée par une mesure de spectrophotométrie à 260 nanomètres, puis une simple division apporte le résultat en masse d'ADN, facilement convertible en paires de bases (environ 1,1 10^{-21}g par paire de bases). A la fin des années soixante, un facteur mille avait été découvert entre la quantité d'ADN contenue dans *Escherischia coli* et celle contenue dans une cellule de mammifère. Dès 1968, des articles pionniers montraient par des expériences de dénaturation/renaturation que près de 50 % de l'ADN des mammifères était constitué de séquences moyennement ou hautement répétées. Le génome des eucaryotes apparaissait dès lors comme une «mer de nucléotides» dans laquelle les gènes faisaient figure de rares «îlots». En conséquence, il semblait clair jusqu'au début des années 1980, que seuls les procaryotes pourraient bénéficier de façon significative des avancées de la biologie moléculaire. En effet, la connaissance des génomes eucaryotes resta marquée par des progrès sporadiques concernant la structure et la fonction d'un nombre limité de gènes. Le tournant primordial, au moins en ce qui concerne la théorie, date sans doute de 1980, quand un article de David Botstein suggéra de saturer le génome humain de marqueurs génétiques, bref, de réaliser une carte génétique de notre espèce (Botstein *et al* 1980). L'enjeu d'une telle carte était évident : identifier les déterminants moléculaires des 3000 maladies génétiques décrites chez l'Homme et comprendre la logique de fonctionnement des gènes sous-tendant la variabilité génétique humaine en dehors de la pathologie. L'audace de la proposition était énorme : à la différence de la drosophile, les polymorphismes à effet visible sont rares dans l'espèce humaine (comme dans les espèces de mammifères d'élevage), ce qui impliquait la nécessité de trouver d'autres variants identifiables à l'échelle des protéines et surtout de l'ADN.

Un des défis majeurs de l'agronomie moderne est l'identification des gènes sous-tendant la variation continue des caractères héréditaires des animaux de rente. La carte génétique constitue la première étape indispensable de ce défi.

1 / Définitions

Etudiant deux mutations chez le pois de senteur au début du siècle, William Bateson et R.C. Punnett découvrirent dans la descendance F2 des proportions s'écartant fortement des proportions mendéliennes attendues : 9/3/3/1. L'explication de ces résultats surprenants dut attendre le développement de la génétique de la drosophile, et l'interprétation par T.H. Morgan de la liaison génétique. Morgan suggéra que l'écart aux proportions attendues résultait de la présence des deux gènes sur le même chromosome. Lors de l'association des chromosomes homologues à la méiose, un échange de matériel (crossing-over ou enjambement) peut se produire entre les chromatides maternelles et paternelles résultant alors en des chromosomes recombinants (figure 1).

L'unité de carte génétique est définie comme la distance entre locus chromosomiques pour lesquelles un produit méiotique sur cent est recombinant. Cette fréquence de recombinaison est appelée centimorgan (cM) et c'est l'unité de distance géné-

Figure 1. *Comportement des chromosomes à la méiose.*

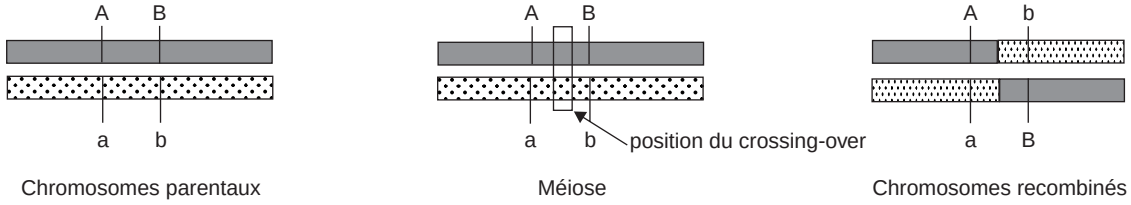
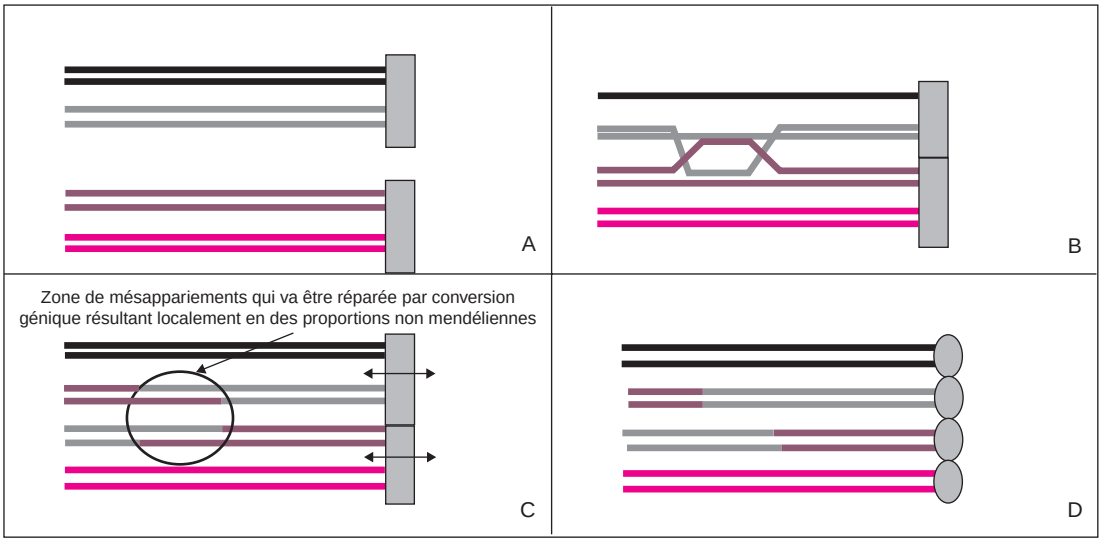


Figure 2. *Détail du crossing-over à la méiose aboutissant à la formation de quatre gamètes dont deux sont recombinés.*



tique.

La cartographie génétique consiste à établir l'ordre et la distance génétique d'une série de locus positionnés physiquement sur le même chromosome. Elle repose donc en définitive sur l'analyse génétique d'individus informatifs, c'est-à-dire hétérozygotes aux locus étudiés. Selon la densité en gènes ou en marqueurs de la région étudiée, on peut avoir recours à trois niveaux d'analyse, liés aux degrés de résolution de ce type de carte : l'analyse de ségrégation, l'analyse d'individus recombinants, et l'analyse du déséquilibre de liaison. Dans tous ces cas, l'outil biologique du généticien est la recombinaison méiotique. Pour cette raison, les termes carte génétique et carte méiotique sont synonymes. Il est frappant que la carte génétique, cet outil si puissant pour localiser des gènes puis les identifier, repose sur un mécanisme méiotique encore mal compris, le crossing-over. Si les raisons biologiques (finalistes) de l'appariement méiotique sont claires en termes de brassage de gènes, la base moléculaire sous-jacente est complexe. La recombinaison génétique est en fait initiée par des cassures double-brin de l'ADN distribuées tout au long des chromosomes et souvent dans des zones fortement 'recombinogènes' (hot-spots de recombinaison). Le microscope électronique a révélé la mise en place d'une structure protéique spécifique de l'appariement, le complexe synaptonémal, résultant de l'association de plusieurs polypeptides formant une structure symétrique tripartite, l'appariement commençant au niveau des télomères chez les animaux. Il est probable que ce complexe soit le berceau des événements de cassure au stade pachytène. Suite aux cassures double brins, des zones d'appariements en double hélice se forment entre les chromatides des chromosomes homologues, permettant les échanges subséquents (figure 2).

La zone d'appariement des deux chromatides aboutissant à la cassure et à l'échange des chromatides, autrefois supposée punctiforme, s'étend en fait sur quelques centaines à quelques milliers de bases. Il est à noter que le polymorphisme naturel de l'ADN implique que les deux segments appariés ne soient pas parfaitement homologues et présentent des bases isolées ou des petites régions simple-brin. Les systèmes de réparation cellulaire (mutS, mutL, système SOS) se chargent d'effacer ces quelques erreurs. Dans la zone impliquée, on assistera donc à des phénomènes de conversion génique qui donneront localement une répartition non mendélienne des produits méiotiques. Bien entendu, par rapport à la longueur du chromosome, ces lissages représentent une faible proportion. Il n'en reste pas moins que ces mécanismes illustrent la complexité réelle des mécanismes méiotiques. L'analyse méiotique globale que l'on effectue pour construire une carte de liaison masque en fait cette complexité sous-jacente. Les écarts souvent observés entre les distances physiques et les distances génétiques, les différences locales du taux de recombinaison le long des chromosomes ou entre les sexes pour différentes espèces ne pourront probablement être expliqués *in fine* que par l'analyse moléculaire.

2 / L'analyse de ségrégation

L'analyse de ségrégation est basée sur l'analyse de la transmission allélique au sein de familles informatives. Pour construire une carte, il est donc nécessaire de disposer de telles familles, d'une part, et de

posséder des marqueurs génétiques d'autre part.

Tout polymorphisme génétique constitue de fait un marqueur génétique. La détection du polymorphisme peut être immédiate (polymorphisme à effet visible) ou nécessiter des systèmes d'analyse plus ou moins sophistiqués. A la fin des années 1970, le généticien japonais Motoo Kimura faisait l'hypothèse qu'une part très importante de la variabilité des protéines consiste en des mutations sans effet sélectif. Ces mutations neutres se retrouvent encore plus fréquemment dans les régions non codantes de l'ADN, majoritaires chez la plupart des vertébrés. La famille de marqueurs les plus utilisés pour la construction de cartes génétiques est celle des microsatellites, en particulier des dinucléotides de type (TG)n. Présents tous les 50 à 100 kilobases d'ADN chez les mammifères, donc nombreux et bien répartis, ces microsatellites présentent en moyenne 6 ou 7 allèles. Leur polymorphisme est, de plus, facilement analysable par des moyens automatiques (séquenceur d'ADN). D'autres types de marqueurs ont également servi de base à la construction de cartes (AFLP , RLGS, RAPD, etc). L'analyse automatique des mutations ponctuelles de l'ADN (ou SNP pour Single Nucleotide Polymorphism) à l'aide de puces à ADN (DNA chips) se développe par ailleurs rapidement et pourrait supplanter les microsatellites pour certaines applications de cartographie fine. Ce type de polymorphisme, quoique moins informatif que la plupart des microsatellites, est présent tous les 500 à 1000 paires de bases dans l'ADN des mammifères, suggérant qu'un microsatellite pourrait aisément être remplacé par 5 ou 10 SNP consécutifs.

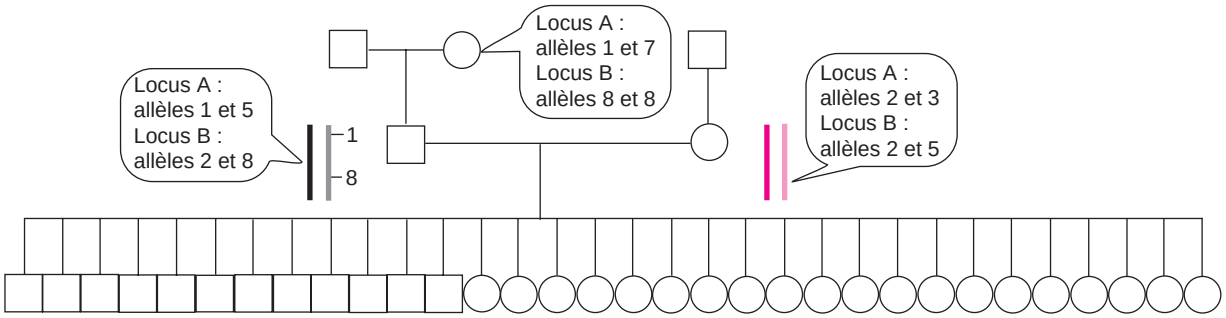
La qualité d'un marqueur génétique peut être évaluée par le PIC (Polymorphic Information Content, Botstein *et al* 1980), qui se calcule ainsi pour un marqueur codominant :

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, p_i^2\, p_j^2$$

Dans cette formule, les p_i correspondent aux différentes fréquences alléliques. On soustrait tout d'abord à l'unité la probabilité de rencontrer un homozygote, nécessairement non informatif. La deuxième somme correspond à la probabilité de l'ensemble des événements pour lesquels la phase ne peut être déduite de la garniture allélique, c'est-à-dire, par exemple, si les deux parents ainsi que leur descendant sont de génotype A_iA_j. En effet, dans ce cas, la probabilité du génotype des parents est $(2p_ip_j)^2$. Dans ces conditions, la probabilité que leur descendant soit A_iA_j est de 1/2, d'où le terme de la deuxième somme $(2p_ip_j)^2$. Une dérivation du PIC montre aisément que la valeur maximale est atteinte quand toutes les fréquences sont égales à 1/n, n étant le nombre d'allèles.

Le deuxième élément indispensable à la construction d'une carte génétique est un ensemble judicieux de familles informatives. Idéalement, une famille idéale pour la cartographie génétique contiendra un grand nombre de descendants de type pleins frères ou pleines sœurs. Dans ce cas en effet, un génotypage donnera, sous réserve d'informativité deux informations méiotiques (mâle et femelle). Idéalement encore, la connaissance du génotype de un ou plusieurs grands-parents permettra de connaître la phase de façon certaine (figure 3). Dans ces conditions, les recombinants et les non recombinants pourront être additionnés à travers plusieurs familles différentes. Dans le cas contraire, l'analyse statistique est menée indépendamment, famille par famille, ce qui diminue les effectifs traités et la puissance de détection des liaisons. Il faut cependant noter que quand les distances génétiques deviennent très faibles, la phase est également connue avec une certitude quasi absolue, et un raisonnement interfamilial devient valide (voir paragraphe 3).

Figure 3. *Exemple d'une famille idéale pour la cartographie génétique. Les descendants sont de nombreux frères et sœurs, ce qui permettra d'obtenir de l'information de la ségrégation paternelle et maternelle. La grand-mère paternelle est de génotype connu pour les deux locus étudiés. Il en résulte que l'on peut déduire de façon certaine l'association physique des allèles chez le père. A1 et B8 sont sur un chromosome, A3 et B2 sur le deuxième. Cette information (la phase) permet d'identifier de façon certaine (non statistique) les recombinants parmi les descendants.*



Enfin, il est particulièrement intéressant dans le cas d'un effort international de cartographie d'un génome, d'utiliser les mêmes familles informatives à travers le monde. Dans ces conditions, les données de typage peuvent être centralisées et interprétées globalement.

Un autre aspect particulièrement important de l'analyse de ségrégation est la gestion des calculs. En première approximation, le taux de recombinaison fournit une distance en centimorgan. Dès les premières études sur la drosophile, on s'est cependant aperçu que les distances génétiques suivaient généralement l'inégalité triangulaire suivante :

$$D(A,C) < D(A,B) + D(B,C)$$

La raison biologique principale est la possibilité de doubles crossing-over qui ne sont pas dénombrés (et qui devraient en fait être comptés deux fois) dans le calcul de distance. Certains de ces crossing-over peuvent ne pas être visibles, mais demeurent détectables quand leur nombre, évalué par le pro-

duit des taux de recombinaison de deux intervalles consécutifs diffère du nombre de doubles recombinants réellement observés ; on parle alors d'interférence. De toutes les façons, l'additivité vraie ne peut être observée que sur de très courtes distances. Pour des distances plus grandes, on peut faire appel à une fonction cartographique :

$m = -\text{Log}(1-2\theta)/2$ (fonction de Haldane 1919)

m représentant la valeur de distance corrigée et θ le taux de recombinaison mesuré. D'autres fonctions de distances ont été élaborées telles celle de Kosambi en 1944 ($m = \text{atanh}(2\theta)/2$) et présentent chacune leurs avantages et inconvénients.

Le test de liaison/non liaison est basé sur une recherche du maximum de vraisemblance, appelé dans ce cas Lod (Logarithm of the odds) score. Le principe est simple, il consiste à estimer le taux de recombinaison et à tester s'il est égal à 50 %. Pour un taux de recombinaison θ, la vraisemblance des observations s'écrit

$L(\theta) = \theta^r(1-\theta)^{n-r}$

où r est le nombre de recombinants et n-r le nombre de non recombinants. Le logarithme décimal du rapport $L(\theta)/L(\theta=0,5)$ doit atteindre la valeur 3 pour être considéré significatif. Bien que cette règle de décision ne repose pas sur la théorie classique des tests d'hypothèse, il est d'usage de considérer que cette valeur seuil correspond à un risque global de 5 % de détecter une liaison inexistante lors de la cartographie d'un génome. En effet, le génome des mammifères est constitué approximativement de 60 segments de 50cM, ce qui multiplie par 60 le risque de détecter une liaison non réelle. Quand la phase n'est pas connue, il est nécessaire d'ajouter un deuxième terme inversant r et n-r puisqu'on ne peut certifier qui sont les recombinants. Pratiquement, on fait varier θ pour trouver l'extrêmum de la courbe de Lod Score, ce qui permet de déterminer une distance de recombinaison la plus probable. Il est clair que ce type de méthode s'applique bien à l'analyse 2-points. Des logiciels ont été développés pour construire des cartes multipoints, également basés sur des méthodes de maximisation de la vraisemblance (logiciels LINKAGE, CRI-MAP, MAP-MAKER). En général, ces programmes fabriquent les cartes par des algorithmes incorporant successivement des marqueurs dans un échafaudage primaire qui s'enrichit progressivement. En effet, le nombre de permutations possibles pour une analyse complète de tous les ordres dépasse rapidement les capacités de calculs des ordinateurs les plus puissants.

Pour les animaux domestiques, l'objectif originel de la cartographie était la constitution d'un réseau d'environ 200 marqueurs régulièrement espacés (tous les 15 à 20 cM environ). Pour ce genre de résolution, et en tenant compte du polymorphisme des marqueurs, 300 méioses informatives suffisent en théorie (Elsen *et al* 1994).

3 / L'analyse d'haplotypes

3.1 / A l'intérieur de familles (inbred)

Pour des distances génétiques très petites (< 1-3 cM), les logiciels d'analyse de liaison ne parviennent plus à ordonner tous les marqueurs de façon non ambiguë. Par exemple, CRI-MAP présente les vraisemblances de position de certains marqueurs à différents endroits de la carte. Une manière de résoudre ce problème consiste à étudier les individus recombinants dans les familles, la position de la recombinaison permettant a priori d'ordonner les marqueurs. Un exemple est donné au tableau 1.

Sur des petites distances, la non connaissance de la phase cesse d'être problématique, et les recombinants peuvent être analysés sur plusieurs familles simultanément (alors que, dans ce cas, les analyses de Lod Score sont indépendantes). Un bon exemple d'application de ce type d'analyse est donné à propos

Tableau 1. *Exemple d'analyse d'haplotypes sur une famille de 47 descendants pour des marqueurs très peu distants (par exemple sur 1 cM). Quatre recombinants sont observés. L'ordre des cinq marqueurs proposé est celui qui minimise le nombre de cassures chromosomiques (crossing-over) chez ces recombinants. Dans de tels cas, les logiciels sont impuissants à positionner de façon fiable les marqueurs par simple analyse de la vraisemblance.*

	M1	M2	M3	M4	M5	Effectif
Type1	8	4	1	3	2	24
Type2	7	5	12	9	4	19
Type3	8	4	12	9	4	1
Type4	7	5	12	3	2	2
Type5	8	5	12	9	4	1

du gène PIS caprin (Cribiu *et al* 2000, cet ouvrage).

3.2 / Hors de structures familiales (outbred)

L'analyse d'haplotypes peut, de façon similaire, être étendue à des analyses non familiales. Ceci est à rapprocher de l'analyse du déséquilibre de liaison (voir paragraphe 4). Chez l'Homme, la population finlandaise constitue un modèle hors du commun des possibilités de la cartographie fine à l'intérieur de certaines populations. Cette population humaine résulte en effet d'un événement fondateur : un nombre limité d'individus est à l'origine de l'ensemble de la population. Parmi ces individus, certains étaient porteurs de maladies génétiques à l'état hétérozygote. La population ayant reçu peu ou pas d'apports d'individus extérieurs, il arrive que les gènes délétères se retrouvent à l'état homozygote parmi leurs descendants éloignés. En conséquence, sur des distances génétiques très courtes, des phases de marqueurs peuvent être conservés (identité par descendance). Du fait que plusieurs dizaines de générations (au moins) se sont écoulés, ces observations permettent une cartographie particulièrement fine, le maintien de la phase impliquant une distance minime. L'étude des recombinaisons parmi les porteurs permet donc de localiser précisément un gène recherché dans une carte très dense.

4 / La cartographie génétique par déséquilibre de liaison

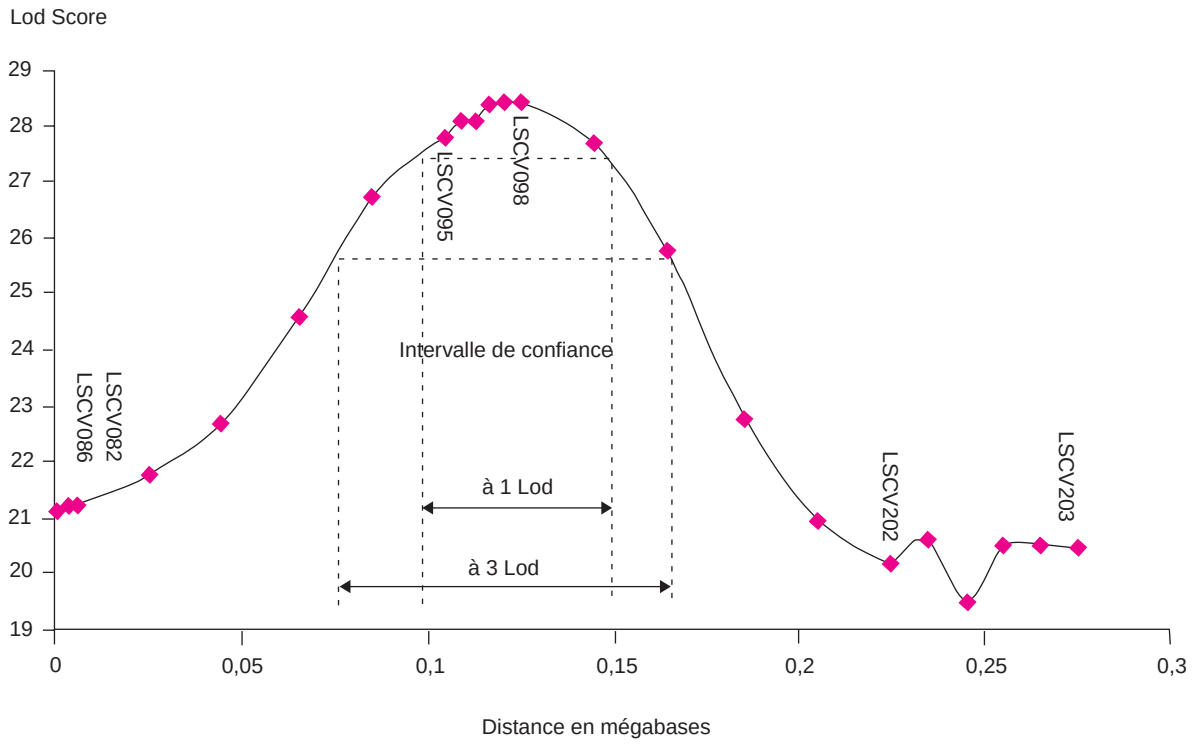
Prenons deux locus bialléliques L1 et L2. Les

allèles de L1 sont A et a, ceux de L2 sont B et b. Si l'on détermine le génotype de 1000 individus quelconques en L1 et L2, on observera quatre possibilités, AB, Ab, aB et ab. Dans le cas général, un quart de chacun des types sera observé. Cette constatation est évidente pour des locus positionnés sur des chromosomes différents. Elle reste exacte quand ils sont positionnés sur le même chromosome, et même dans la plupart des cas s'ils sont relativement proches. Cependant, dans le génome humain, des écarts à ces proportions s'observent dans 5 % des cas pour des marqueurs microsatellites espacés de 4 cM (Huttley 1999). Quand on s'intéresse à des distances plus faibles, les écarts deviennent le cas général. Ces déviations constituent le déséquilibre de liaison. L'association préférentielle d'allèles provient du fait que même dans une population d'individus apparemment non apparentés, le nombre de géniteurs originels est en fait limité. La détection d'un déséquilibre est donc fonction de la distance entre deux marqueurs (ou entre un gène et un marqueur). Les individus pour lesquels on réalise l'observation ont en fait un ancêtre commun séparé de plusieurs générations, et on a perdu la trace réelle de leur origine. A chaque génération, la recombinaison génétique estompe la liaison, ce qui implique qu'à terme, le déséquilibre doit finir par disparaître. Grossièrement, le nombre d'individus non recombinants observés au bout de n générations vaudra $(1- \theta)^n$.

Il apparaît clairement que pour des distances très faibles, ce nombre ne décroîtra que lentement. En outre les événements fondateurs (des sélections d'individus de populations fermées) peuvent régénérer de nouveaux déséquilibres de liaison.

En général, les études de déséquilibre de liaison cherchent à positionner un gène dans une carte de marqueurs préétablie. Elles permettent cependant aussi de positionner relativement des marqueurs. Des modèles de maximum de vraisemblance ont été développés pour les marqueurs pluriallèliques de type microsatellite, en particulier par Terwilliger

Figure 4. *Déséquilibre de liaison sur 200 kilobases dans la région du gène 'sans corne' caprin. Les marqueurs microsatellites sont appelés LSCV. Un pic de LodScore est observé près de LSCV098 et LSCV095. L'étude des haplotypes permet de plus de positionner la mutation causale à gauche de LSCV095. L'intervalle de confiance à une unité de LodScore restreint la position du gène à une cinquantaine de kilobases. La courbe est obtenue à l'aide du programme DISMULT (Terwilliger 1995).*



Conclusion

L'analyse génétique fournit un ensemble d'outils cartographiques pour des distances s'échelonnant de 10 cM à 0,01 cM. Il convient d'insister encore une fois sur la différence entre les analyses familiales, portant sur la ségrégation entre deux générations consécutives, et les analyses populationnelles, portant sur de nombreuses générations et permettant en théorie une cartographie beaucoup plus fine. Les marqueurs polymorphes innombrables que constituent les SNP suggèrent la possibilité de se rapprocher à volonté des gènes que l'on désire cloner. La cartographie génétique utilisant les marqueurs moléculaires éloigne chaque jour de l'obsolescence les techniques développées par Morgan au début de ce siècle.

Références

Botstein D., White R.L., Skolnick M., Davis R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics, 32, 314-331.

Cribiu E.P., Schibler L., Vaiman D., 2000. Cartographie fine de la région du gène PIS de la chèvre. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 141-144.

Elsen J.M., Mangin B., Goffinet B., Chevalet C., 1994. Optimal structure of protocol design for building genetic linkage maps in livestock. Theoretical Applied Genetics, 88, 129-134.

Haldane J.B.S., 1919. The combination of linkage values and the calculation of distance between the loci of linkage factors. Journal of Genetics, 8, 299-309.

Huttley G.A., Smith M.W., Carrington M., O'Brien S.J., 1999. A scan for linkage disequilibrium across the human genome. Genetics, 152, 1711-22.

Kosambi D.B., 1944. The estimation of map distances from recombination values. Annals of Eugenics, 12, 172-175.

Terwilliger J.D., 1995 A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. American Journal of Human Genetics, 56, 777-87.