# Travaux dirigés marqueurs et cartographie

T. Schiex, INRA

October 4, 2012

## 1 Cartes haute densité

- On a deux marqueurs. Pour chaque individu, la probabilité d'observer un génotype différent sur ces 2 marqueurs est de $r$ (probabilité de recombinaison). Ici, cette probabilité est de $1/(n-1)$ (car $n$ est grand et à petite distance $r \approx d$).

  Pour que deux marqueurs aient des génotypes identiques, il faut qu'il n'y ait pas recombinaison (probabilité $1-1/(n-1)$) dans tous les cas, qui sont indépendants (entre individus). Donc une probabilité de $(1-1/(n-1))^p$ d'avoir l'identité complète.

  Pour $n = 1000, p = 100$, on a $(1-1/999)^{100} = 0,904701491\ldots$. Il y a donc une forte probabilité d'avoir 2 marqueurs adjacents confondus.

- Sans interférence, la probabilité de recombinaison est indépendante entre les différents intervalles. A chaque paire de marqueurs adjacents qui correspond à deux marqueurs identiques, on perd un marqueur. Sur $n-1$ paires, le nombre moyen de pertes de marqueurs sera donc $n-1$.(probabilité d'identité), soit $(n-1).(1-1/(n-1))^p$.

  Pour $n = 1000p = 100$, on perdra $999 \times 0,90479$ soit 903 marqueurs. Il reste une centaine de marqueurs avec de l'information associée (il faut l'espérer).

- on veut que $(n-1).(1-1/(n-1))^p = n.\varepsilon$. En passant au logarithme, le résultat obtenu est $p = (\log(\varepsilon) + \log(n/n-1))/\log(n-2/n-1)$.

  Pour $n = 100$, $p = 225$, pour $n = 1\ 000$, $p = 2\ 298$, et pour $n = 10\ 000$, $p = 23\ 021$.

- On a $n'!/2$ ordres et $2^{pn*}$ jeux de données possibles avec $n* = n-(n-1).(1-1/(n-1))^p$. Et $n'!/2 \approx 1/2.\sqrt{2\pi n'}(n'/e)^{n'}$. On veut l'égalité comme limite, on passe au logarithme.

  $\log(\sqrt{2\pi n'}/2) + n'\log(n'/e) = pn*.\log(2)$

  Or $pn* = p(n-(n-1)(1-1/(n-1))^p) \approx p(n-(n-1).(1-p/(n-1))) = p(1+p)$ qui ne dépend pas de $n$ en première approximation pour $n$ grand. Et donc on veut

$p(p+1) = (\log(1/2.\sqrt{2\pi n'}) + n' \log(n'/e)/\log(2)$

On approxime $p(p+1)$ par $p^2$ et on a

$p = \sqrt{\log(1/2.\sqrt{2\pi n'}) + n' \log(n'/e)/\log(2)}$

Pour $n = 1\ 000$, $p = 92$ (ce n'est pas une contrainte très informative car en général on a plus de 96 individus). Pour $n = 10\ 000$, $p = 344$ est un peu plus intéressant car c'est un effectif assez rarement atteint. Mais il s'agit de minorants naïfs. La valeur réelle est, sans doute, très supérieure.

# 2 LOD et distorsion

- le marqueur M4 a 34 : 16 comme équilibre allélique. Alors que 25/25 est attendu (ségrégation en 1:1). On calcule un $\Xi^2$ et on obtient $18^2/50 = 6.5$. On a bien un marqueur distordu car le seuil avec 1 degré de liberté à 5% est de $3,84$. On rejette M4.
- Pour estimer le taux de recombinaison en 2 points, on utilise l'estimateur direct, qui est aussi de maximum de vraisemblance $R/(R+NR)$. Ici
  1. M1/2 : R=4, NR=46. $\hat{r} = 0.08$, soit $8,7cm$ Haldane soit $8,1cm$ Kosambi.
  2. M1/5 : R=18, NR = 32. $\hat{r} = 0,36$ soit $71,4cm$ Haldane ou $49,8cm$ Kosambi.
- Pour M1/2, on obtient un LOD de 9: liés selon les critères usuels
- Pour M1/5, on obtient un LOD de $0,6$: pas assez pour conclure à la liaison.

  On en peut pas conclure que M1 et M5 ne sont pas dans le même groupe de liaison pour autant. Il se peut que M2 soit lié à M3 et M3 à M5.

# 3 Vraisemblance et RH

- Les configurations possibles:
  - 00 possible
  - 01 obligée
  - 10 obligée
  - 11 possible
- C'est tout simplement $O_{ij} = \sum_n |M_{ik} - M - jk|$

- On va chercher un ordre minimisant le nombre de cassures obligatoires. On tente de mettre, autant que possible, les marqueurs ayant peu de cassures obligatoires (OCB) entre eux proches les uns des autres. C'est cohérent avec l'idée que plus 2 marqueurs sont proches, moins il y a de cassures. La cassure est rare et l'on utilise un critère de parcimonie (rasoir d'Ockham).
- on modélise le problème comme un problème de voyageur de commerce en utilisant $d_{ij} = O_{ij}$

## 3.1 Probabilités

On note $p$ le taux de rétention et $q = 1 - p$. On suppose $p$ connu et égal à (nombre de 1)/(nombre de 0+nombre de 1).

- examinons chacun des cas:
  - 00: (pas retenu, cassure, pas retenu) ou (pas cassure,pas retenu). Exclusifs donc probabilité: $q((1-c)+cq) = q(1-c+c-cp) = q(1-cp)$
  - 01: (pas retenu, cassé, retenu), donc $qcp$.
  - 10: (retenu, cassé, pas retenu) don $pcq$ encore.
  - 11: (retenu cassé retenu) ou (pas cassé, retenu), donc $p(c.p+(1-c)) = p(c(1-q)+1-c) = p(c-qc+1-c) = p(1-qc)$
- les hybrides sont tous indépendants, on peut donc faire le produit de toutes les probabilités sur chaque hybride. On aura un produit avec 3 types de termes:
  - $pcq$ pour O1 et 10
  - $q(1-cp)$ pour 00
  - $p(1-qc)$ pour 11
- Si l'on note $n_{01}, n_{10}, n_{11} et n_{00}$ les différents comptages dans les deux lignes de génotypes des 2 marqueurs. On a donc:

  $$L(D|p,c) = pcq^(n_{01} + n_{10}).(q(1-cp))_{00}^{n}.(p(1-qc))_{11}^{n}$$

  Pour trouver l'estimateur de vraisemblance maximum, il faut que l'on ait un maximum donc que la dérivée de $L$ s'annule ou, de façon équivalente la dérivée de son logarithme s'annule (plus simple).

  On a $\log(L) = (n_{01}+n_{10}).\log(pcq)+n_{00}.\log(q(1-cp))+n_{11}.\log(p(1-qc))$

  La dérivée par rapport à $c$ s'écrit

  $(n_{01}+n_{10}).(1/c) + n_{00}.(-qp/(q(1-cp))) + n_{11}.(-qp/(p(1-qc)))$

  et elle doit s'annuler. En réduisant au même dénominateur, on va faire apparaître des termes du second degré et donc devoir résoudre une équation du second degré, qui possède deux racines en général.

- on choisira la racine qui est cohérente avec un estimateur $c$ nul si $n_{01} + n_{10} = 0$.